Bioinformatics Methods and Protocols

生命科学实验指南系列

生物信息学方法指南

〔加〕S. 米塞诺 〔美〕S.A. 克拉维茨 著

欧阳红生 阮承迈 李慎涛 等译



科学出版社
www.sciencep.com

(Q-1498.0101)

Bioinformatics
Methods and Protocols

生物信息学

方法指南



生命科学编辑部
联系电话：010-64012501
http://www.lifescience.com.cn
e-mail:info@lifescience.com.cn

ISBN 7-03-014465-1
定 价：58.00 元

销售分类建议：生物医学/生物技术/分子生物学

生命科学实验指南系列

生物信息学方法指南

〔加〕S. 米塞诺　〔美〕S. A. 克拉维茨　著

欧阳红生　阮承迈 李慎涛等　译

科学出版社

北京

图字：01-2003-0476

内 容 简 介

计算机在分析生物学日益增长的海量数据方面起到了不可估量的作用，并推进了现代生物学的快速发展。本书详细介绍了一些重要生物学软件和数据库的使用，同时提供了一些实用的技巧和最新研究进展。全书分为五部分，包括序列分析软件包、分子生物学软件、网络信息资源、计算机和分子生物学的关系、生物信息学教学与最新文献跟踪。内容全面，实用性较强，可帮助生物信息学人员对该学科有更深入地了解。

本书可作为大专院校、科研机构的分子生物学、生物信息学等相关专业的研究生、科研和教学人员的参考书。

The original English language work has been published
by HUMANA PRESS Totowa, New Jersey, U.S.A.
©1999 by Humana Press.
All rights reserved.

图书在版编目(CIP)数据

生物信息学方法指南/欧阳红生，阮承迈，李慎涛等译. —北京：科学出版社，2005
(生命科学实验指南系列)
ISBN 7-03-014465-1

I. 生… II. ①欧… ②阮… ③李… III. 生物信息论 IV. Q811.4

中国版本图书馆 CIP 数据核字(2004)第 114450 号

责任编辑：马学海 庞在堂 彭克里 席 慧/责任校对：朱光光
责任印制：钱玉芬/封面设计：王 浩

科学出版社 出版
北京东黄城根北街16号
邮政编码：100717
http://www.sciencep.com
源海印刷有限责任公司印刷
排版制作：科学出版社编务公司
科学出版社发行 各地新华书店经销
*
2005年2月第 一 版 开本：B5 (720×1000)
2005年2月第一次印刷 印张：26
印数：1—3 000 字数：517 000

定价：58.00 元

(如有印装质量问题，我社负责调换〈环伟〉)

前　言

计算机已成为现代生物学的一个基本组成部分，它帮助管理大量的并日益增长的生物学资料，并在发现新的生物学关系方面继续起着综合的作用，这种基于计算机的生物学方法帮助重塑了现代生物学。我们正置身于生物学革命中，每个科学家都必须提高和训练当今的生物信息学技能，但达到初级水平即可。《生物信息学方法指南》就是要满足这种挑战，并向富有经验的用户提供实用的技巧和目前最新进展的概观。本书以 1994 年出版的两卷套的《序列数据的计算机分析》为基础编著。我们把《生物信息学方法指南》分成五部分，包括在大多数机构都可得到的基本序列分析软件包的全面综述，也有基础入门级的生物信息学课程的设计和内容。此外，本书还介绍了非商业化的专业软件、数据库及因特网上的其他资源，以及针对生物学家现在面临的计算挑战及将来可能的解决方案的讨论。

第一部分，序列分析软件包，介绍目前可得到的一些分析软件包的资源指南，包括在大多数机构可见到的客户服务器 GCG 软件包，几个基于 PC 机和 Mac 机(Macintosh)的适于独立计算的软件包。Staden 软件包也很有特色，因为它是最广泛使用的整套序列分析和装配的软件工具，且针对学术研究可以免费得到。这里也介绍了免费软件的使用，这些软件用于建立一些解决特定需求的分析方案。

第二部分，分子生物学软件，收集了用于完成一些基本生物信息学任务的软件资源。本部分从目前各种计算机平台可得到的免费软件的概观开始，接着是一些特定的例子，其中有用 FASTA、CLUSTAL 多重序列比对进行的序列相似性搜索和种系发生分析。之后，讨论了 Genotator，一个功能非常强大的序列注释和展示套件，它整合了多种不同分析输出适于出版的格式。本部分最后讨论了常见图像分析技术。

第三部分，网络信息资源，主要介绍了基本的一级(primary)序列数据库和各种可得到的分析工具。本部分也有对临床资源进行的独特描述，临床资源正快速成为分子医学新兴领域的整体部分。一级序列分析方法有用 MatInspector 鉴定转录控制区的手段，也有对目前基因鉴定方法的评述。最后讨论了寡聚核苷酸和 PCR 引物的设计及通过万维网分配实验方法和试剂的非常实用的模式。

第四部分，计算机和分子生物学，作者直接阐述了基于计算机分析的局限性和可能存在的解决办法。本部分最后提出了仍然不能回答的问题，即怎样从 A、C、G 和 T 序列串中检测有生物学意义的模式。

生物信息学教学很快成为大多数大学核心课程的组成部分。本书第五部分，作者推荐了生物信息学入门课程的设计。本部分深入分析了如何跟踪日益增加的

文献。简而言之，为在现代生物学成功生涯中日常遇到的问题提供了基本的、实用的答案。

生物信息学为目前的现代生物学革命成为可能提供了帮助。只有了解并明智地使用这些资源，我们才能向前推进。在本书中，对每个学科领域广泛的了解主要是为了帮助那些刚刚开始用计算工具诠释生物学问题的人把握方向。我们相信，在这个独特的软件集和解释例子的指导下，即使初学者也能很快应付每个计算问题，并获得满意的结果。

Stephen A. Krawetz
Stephen Misener

(欧阳红生　译)

编 写 成 员

Ashok Aiyar • *University of Wisconsin-Madison, Madison, WI*
Roger Anderson • *Anderson Unicom Group, Inc., Yorba Linda, CA*
Kathryn F. Beal • *MRC Laboratory of Molecular Biology,Cambridge, UK*
James K. Bonfield • *MRC Laboratory of Molecular Biology, Cambridge, UK*
Timothy G. Burland • *DNASTAR, Madison, WI*
Brian Fristensky • *University of Manitoba, Winnipeg, Manitoba, Canada*
Don Gilbert • *Indiana University, Bloomington, IN*
Nomi L. Harris • *Lawrence Berkeley National Laboratory, Berkeley, CA*
Jack P. Jenuth • *Base4 Bioinformatics, Mississauga, Ontario, Canada*
Lila Kari • *University of Western Ontario, London, Ontario, Canada*
Jeffrey A. Kramer • *Monsanto Life Science Company, St. Louis, MO*
Stephen A. Krawetz • *Wayne State University School of Medicine, Detroit, MI*
Maryann Labant • *Anderson Unicom Group, Inc., Yorba Linda, CA*
Laura F. Landweber • *Princeton University, Princeton, NJ*
Avi Orr-Urtreger • *Genetic Institute, Tel Aviv, Israel*
William R. Pearson • *University of Virginia, Charlottesville, VA*
Promila A. Rastogi • *Oxford Molecular Group, Campbell, CA*
Keir Reavie • *Wayne State University, Detroit, MI*
Jeffry A. Reidler • *Scion Corporation, Frederick MD*
Jacques D. Retief • *University of Virginia, Charlottesville, VA*
Patricia Rodriguez-Tomé • *EMBL European Bioinformatics Institute, Hinxton, Cambridge, UK*
Steve Rozen • *Whitehead Institute for Biomedical Research, Cambridge, MA*
Helen Skaletsky • *Whitehead Institute for Biomedical Research, Cambridge, MA*
Gautam B. Singh • *Oakland University, Rochester, MI*
Rodger Staden • *MRC Laboratory of Molecular Biology, Cambridge, UK*
Paul Stothard • *University of Alberta, Edmonton, Alberta, Canada*
Gary H. Van Domselaar • *University of Alberta, Edmonton, Alberta, Canada*
Thomas Werner • *Institute of Mammalian Genetics, Neuherberg, Germany*
David S. Wishart • *University of Alberta, Edmonton, Alberta, Canada*
David D. Womble • *Wayne State University School of Medicine, Detroit, MI*
Yuval Yaron • *Genetic Institute, Tel Aviv, Israel*

目　录

第一部分　序列分析软件包

1 GCG：序列分析程序威斯康星软件包

David D. Womble

1.1 引言

GCG 程序，又称为“威斯康星软件包”，是具有强大功能的操作、分析和比较核苷酸和蛋白质序列的整套软件工具[1]。GCG 是遗传学计算机小组(Genetics Computer Group)的缩写，该小组隶属于牛津分子小组(加州坎贝尔)。威斯康星软件包含有 130 多个程序，每个程序都可以作为完成特定任务的工具，例如，翻译核苷酸的编码序列、分析限制酶切位点。大多数 GCG 程序用文件的方式输入数据，并将分析结果写到另一个文件中。很多 GCG 程序的输出文件可作为其他 GCG(或另一些软件包)程序的数据输入文件。很多情况下，复杂的问题需通过连续使用几个 GCG 程序得到解决。

威斯康星软件包通常安装在网络上的共享计算机上，如安装在含 UNIX 操作系统的服务器上，这样，用户可在远程终端上通过自己的个人计算机或其他终端访问 GCG 程序。有几种不同的方法运行 GCG 程序，软件包中包括了两种方法：一种是命令行界面，它是一种传统方法，用户键入一个 GCG 程序名开始交互式程序应用；另一种是图形用户界面(Graphical User Interface，GUI)，称为 SeqLab。此界面中，用户打开一套 GCG 程序的窗口，采用图形交互式方式选择序列和程序功能。SeqLab 也包括一个功能强大的用不同色彩作标记的图形界面用户序列编辑器。但在每一种界面中，所有程序操作方式都相似。用户一旦熟悉怎样运行软件包中的一个程序，所有的其他程序都能用同样的模式运行。根据作者的经验，刚开始接触 GCG 程序的学生常用易于使用的 SeqLab 图形界面，而有经验的 GCG 用户常用命令行，因为采用命令行运行更快捷，特别是在远程终端上通过网络运行更是如此。这两种界面都能很好运行。最近引入的基于网页的界面，称为 SeqWeb，也可以从 GCG 得到。这种界面允许用户通过 Netscape Communicator 和 Internet Explore 等网页浏览器运行 GCG 程序和操作序列文件。GCG 软件包基于互联网网页的界面见第 2 章。

1.2 材料

本章描述的方法是基于安装在与 TCP/IP 网络相连的 UNIX 操作系统共享计算机上的第 9.1 版 GCG 程序软件包[2]。该软件包能安装在几种不同的计算机系统上，如运行数字 UNIX4.0 的 Digital Alpha 机、运行 6.2、6.3 或 6.4 版 IRIX 的基于 RISC 的 Silicon Graphics 机和运行 2.51 或 2.6 版 Solaris 的基于 SPARC 的 Sun 机。该软件包也可运行在以 6.2 版 OpenVMS 为运行环境的 Digital Alpha 机上。安装维护含全套数据库的威斯康星软件包至少需要 15G 的硬盘空间。随着数据库的扩展，所需硬盘空间需要快速增加。个人用户文件也需要额外的硬盘空间。建议至少应有 128M 核心内存和 200M 的虚拟内存。程序通常运行于 UNIX 环境中的 C 环境中。软件包可以从遗传学计算机小组得到，它们的地址是：3575 Science Drive, Madison, WI 53711, 电话:(608)231-5200, 传真:(608)231-5202, 电子邮件地址：info@gcg.com，网站地址：http://www.gcg.com。

GCG 程序能在 UNIX 计算机控制台(console)或远程工作站(即运行 Windows 或 MacOS 个人计算机)上直接操作。如果使用命令行操作程序，应该使用含 VT100 终端仿真器的运行远程登录软件的终端或 PC 机。如果使用 SeqLab，应该使用 X-Windows 终端或运行 X-Windows 服务器软件的个人计算机。

大多数 GCG 程序的结果保存为常见的文本(ASCII)文件。文本文件可以使用任何文本和文字编辑器进行进一步的操作。此外，很多 GCG 程序的结果以图形方式输出，如限制性内切核酸酶图谱、RNA 二级结构预测。图形输出的方式有：在终端屏幕上显示、在与终端相连的打印机或绘图仪上打印或保存成文件用于以后显示或打印。为了在屏幕上显示图形，需要图形终端或仿真器。X-Windows 仿真器可在屏幕上显示图形，也可用 SeqLab 图形界面。GCG 程序包所带的说明指出各种终端和图形软件都适用于本软件包。作者的建议见 1.4 节。

要打印 GCG 图形，PostScript 或 HPGL 图形语言的机器都可使用。为了在用户终端的打印机上直接从远程服务器上打印图形，需要一个使打印处于透明方式(transparent mode)的终端程序，以便文件直接从打印机输出，而无需个人计算机处理。

1.3 方法

1.3.1 程序描述

威斯康星软件包中共有 130 多个程序。尽管各个程序都能作为独立的工具使

用，但为了便于描述，这些程序可根据功能进行分组。本节介绍软件包中一些通用程序的功能，同时对某些程序进行简要描述，并含有一些例子。尽管对 GCG 程序的完整介绍不是本章的范围，但这些例子可提供足够的信息，使读者了解本软件包中的工具箱的总体内容。

1.3.1.1 比较

1) 配对比较

这些程序可以将一个序列与第二个序列进行比较。可选项有生成两个序列最优的全局(global)比对，找到两个序列的最相似的片段(bestfit)，或者形成序列相似性的 X/Y 图(compare/dotplot)。

2) 多重比较

PileUp 程序采用渐进和配对比对(pairwise alignment)的方法对多组相关序列进行多重序列比对分析。本组中的另一些程序(SeqLab)可手动编辑比对的序列，显示比对序列的各种属性或从比对序列生成用于数据库检索的流程。

1.3.1.2 数据库检索

1) 文献检索

LookUp、StringSearch 程序可通过名称、登录号、作者以及其他关键词查找序列。

2) 序列分析

在这些程序组(BLAST、NetBLAST、FAST 等)中的程序可以在数据库中检索与待查序列相似的序列。NetBLAST 可直接检索美国国家生物技术信息中心(NCBI)的数据库，其他程序可检索本地安装的数据库。

1.3.1.3 编辑和发表

该组程序中有编辑单个序列文件的程序(SeqEd)，也有编辑多个序列文件的程序(LineUp、SeqLab)，同时也可对将要发表的序列数据或质粒图谱作准备。

1.3.1.4 进化关系分析

程序 PAUPSearch、PAUPDisplay、Distances、GrowTree 及 Diverge 可以进行多重比对比较，分析序列的相似性和进化关系。

1.3.1.5 片段组装

GCG 片段组装系统是一套将测序项目得到的序列数据组装成连续序列的程序。

1.3.1.6 基因查找和模式识别

此组程序超过 12 个，有 TestCode、Frames、Motifs 等，这些程序可以帮助识别蛋白质的编码区、蛋白质的结合基序、直接的重复、其他模式以及其他类似的任务。

1.3.1.7 输入和导出

该组有 15 个这种程序，可辅助输入序列数据和对各种格式的序列文件进行格式转换，可转换的格式有 GCG、Staden、EMBL、GenBank、IntelliGenetics、PIR 和 FASTA。

1.3.1.8 绘图

绘图程序(Map、MapPlot、MapSort 等)生成和显示限制酶切图、可读框图、肽消化图、T1 核酸酶消化图、质粒图等。

1.3.1.9 引物挑选

Prime 程序挑选用于聚合酶链反应(PCR)实验和 DNA 测序所用的寡聚核苷酸引物。

1.3.1.10 蛋白质分析

蛋白质分析程序(PeptideMap、PepPlot、PeptideStructure 等)能辅助确定蛋白质氨基酸序列有关的信息，如确定等电点、定位功能基序、预测蛋白质二级结构、分析抗原性质和分泌信号。

1.3.1.11 RNA 二级结构

该组程序(Mfold、StemLoop 等)能按 Zuker 法[3]及确定反向重复序列位置的方法预测 RNA 二级结构并以多种格式显示 RNA 二级结构。

1.3.1.12 翻译

翻译程序(Translate、BackTranslate、PepData 等)将核苷酸序列翻译成肽序列或进行相反的工作。

1.3.1.13　工具程序

1）序列工具

该部分有几个实用程序(Reverse、Shuffle、Simlify 等)，功能有生成反向的核苷酸序列、随机化序列或用 X 字母取代低复杂度区序列。

2）数据库工具

用这些程序，可以从任何 GCG 格式的序列中生成 GCG 个人数据库，将任何 GCG 序列连接到一个数据库中，所形成的数据库可以被 Blast 检索，也可从序列中随机提取序列片段。

3）打印和绘图工具

这些程序(Lprint、ListFile 等)用于显示、打印和绘制 GCG 结果文件，将文本文件或图形文件连接到各种显示、打印或绘图装置。更多的显示或打印 GCG 结果文件的信息见 1.3.5 节。

4）文件和其他小工具

许多其他小工具程序(ChopUp、Replace、Reformat 等)辅助操作文本文件，打印 GCG 文件及其他任务。

1.3.2　数据库

威斯康星软件包中含有一套综合序列数据库。其中有 GenBank 和 EMBL 核酸序列数据库(EMBL 数据库有删节，以避免与 GenBank 重复)、PIR 和 SwissProt 蛋白质序列数据库。数据库中的序列是 GCG 文件格式的，因此它们能直接作为 GCG 程序的输入文件。因为大多数序列存在于数据库中，因此每个用户无须自行搜集这些序列的拷贝，只要查阅数据库中的拷贝就可以搜集到序列数据。还包括了 GCG 程序所用的各种数据库，如限制酶、分值矩阵(scoring matrix)、水解酶及试剂、蛋白质分析数据文件、转录因子数据库(TFD)、密码子使用频度表、翻译表和蛋白质位点和模式字典 PROSITE。这些数据以文本文件形式保存，个人可以按照自己的需要或特定目的进行检索和编辑。

1.3.3　界面

威斯康星软件包有两种界面：命令行界面和名为 SeqLab 的图形用户界面。用命令行界面，用户键入一个 GCG 程序名称就开始一个程序的交互式对话。然后，提示用户运行程序所需的信息，如输入序列文件的名称、让用户从各种备选项的菜单中选择程序怎样操作。在最后一次按回车键后，程序运行。运行后通常将结果存放到一个文件中。所有的 GCG 程序从命令行的操作相似。因此，一旦熟悉一个程序的操作过程，可以用所熟悉的方式操作其他程序。从命令行操作程序也

可脚本化运行，脚本化运行时含命令行开关，这是用多个输入文件多次运行 GCG 程序的高效运行方法。为了从远程终端中使用命令行界面，需要将终端模拟程序经远程登录，与 GCG 服务器相连。终端模拟程序使用 VT100 终端功能。图 1.1 举例示意了命令行界面。

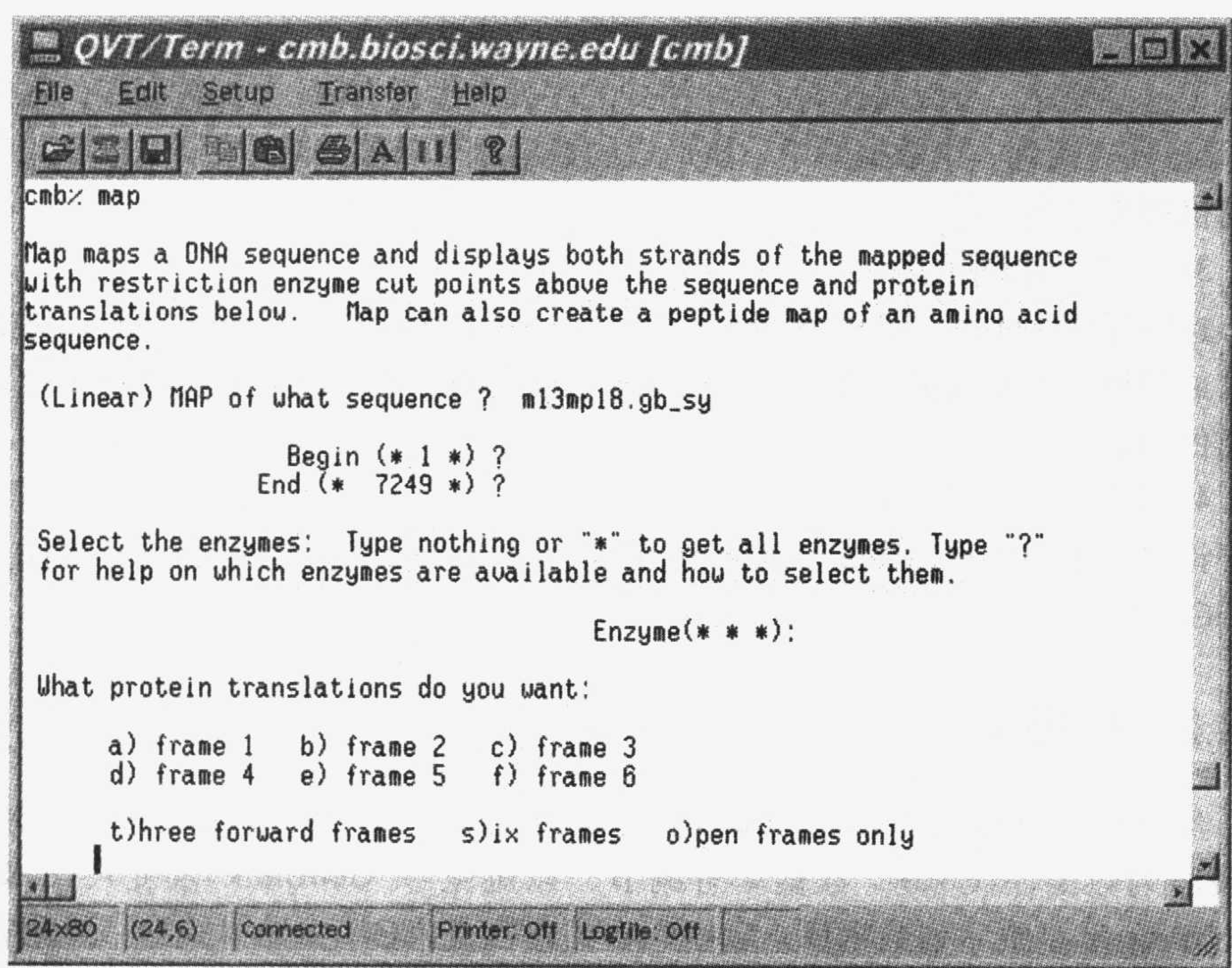


图 1.1 GCG 命令行界面

SeqLab 是 GCG 图形用户界面，它提供了操作威斯康星软件包更方便的使用方法。使用 SeqLab 的下拉菜单可选择程序对序列进行操作。当从下拉菜单中选择一个 GCG 程序时，出现程序专用的一个独立的窗口。然后用鼠标点击选项，可选项有分析哪个序列，接着按 Run 按钮。所选的 GCG 程序的结果列在另一个称为 Output Manager Window(输出管理窗口)中。然而 SeqLab 程序功能超过命令行界面程序，对各碱基或残基或已知的序列性质有更丰富的视觉显示。这种视觉显示使得手动编辑序列或生成和处理多重序列比对更加容易。SeqLab 所用的图形用户界面称为 X-Windows，它是运行 UNIX 操作系统计算机的一种窗口系统。使用 SeqLab，需要一种 X-Windows 显示，如运行在 Windows PC 机或 Mac 机上的 X 服务器程序，或运行 X-Windows 的工作站。图 1.2 显示的是 SeqLab 界面的一个例子。

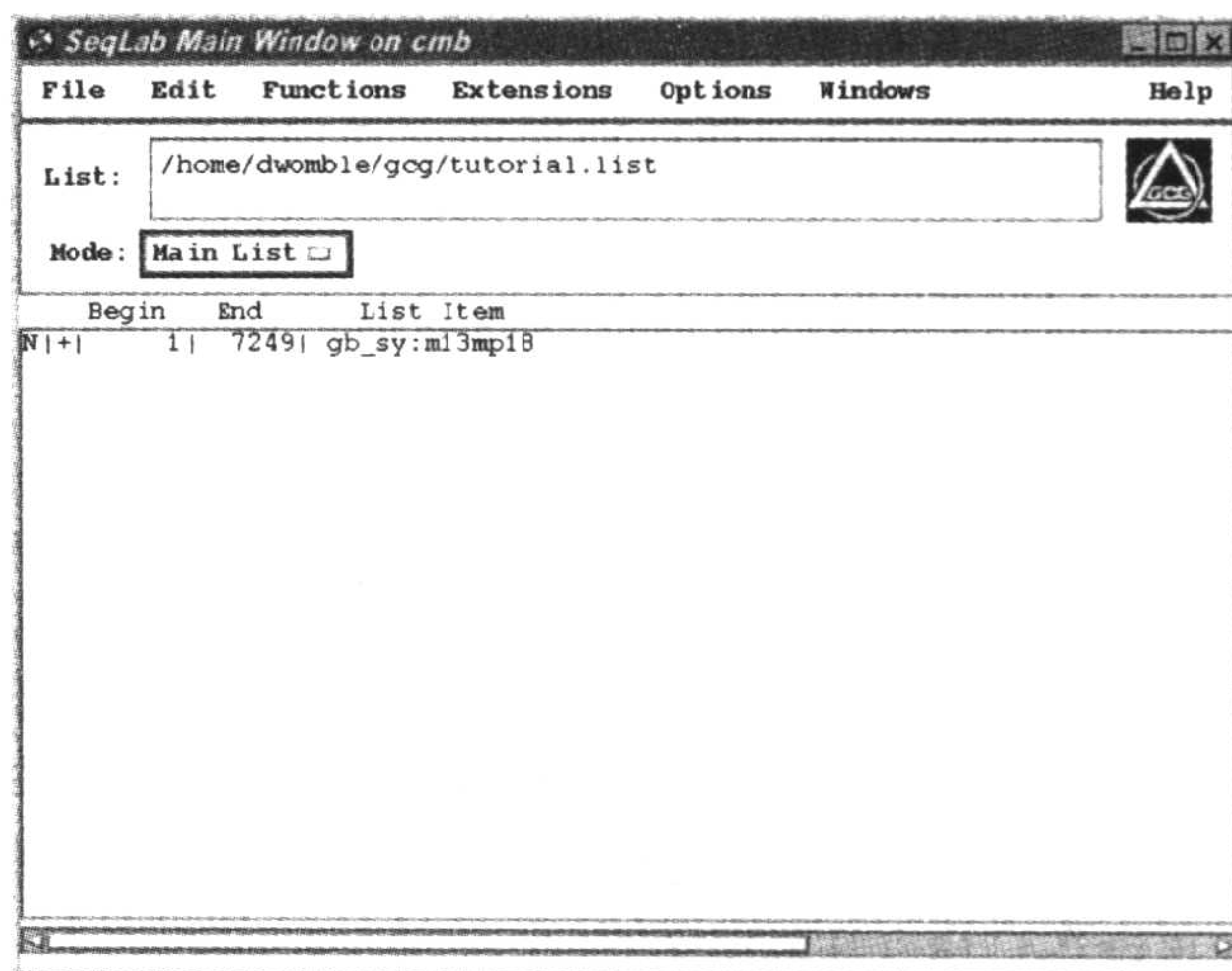


图 1.2　GCG SeqLab X-Windows 界面

一个名为 HYGCGmenu 的程序(GCG 超文本菜单)能用于增强命令行界面的功能。HYGCGmenu 生成一套 GCG 程序屏幕菜单，这样可以用"点击"的方法操作 GCG 程序。箭头指针键被用于选择一个序列并启动交互式 GCG 程序对话。在 HYGCGmenu 中，GCG 程序按功能组织成菜单，这样不用记住各个 GCG 程序的名称也能容易选定程序。HYGCGmenu 也有一套目录浏览和文件管理工具，如拷贝、改名、编辑等，以增强其效能。用 HYGCGmenu 的一个好处是在用户端不需要额外的软件，通过远程登录终端的 VT100 模拟器进行操作。HYGCGmenu 不由 GCG 生产，也不由威斯康星软件包提供，但能被各教育用户或系统管理员免费下载(见 1.4 节)。图 1.3 显示的是 HYGCGmenu 的一个例子。

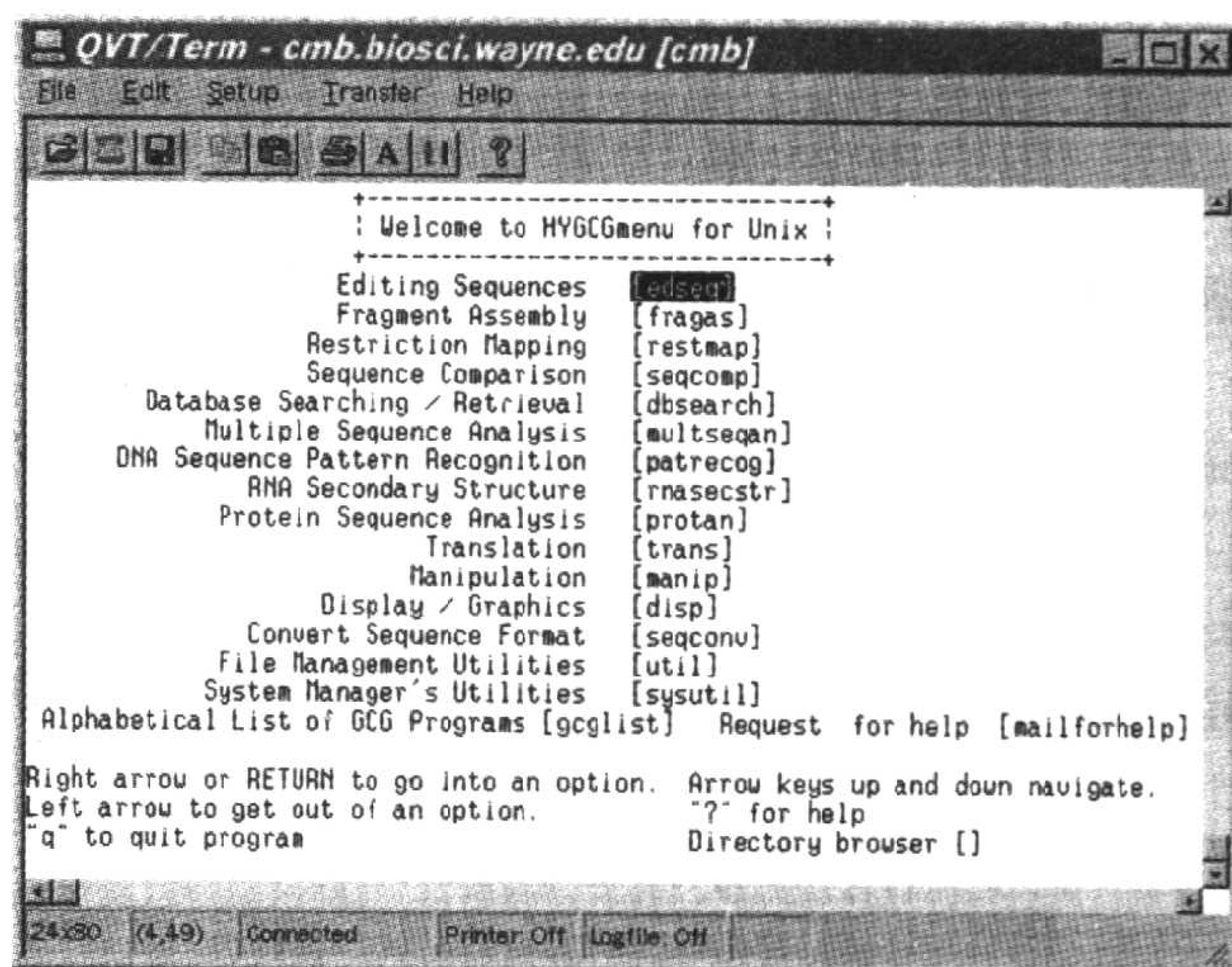


图 1.3　GCG 的超文本菜单 HYGCGmenu

另一个运行 GCG 程序的界面名为 SeqWeb。SeqWeb 作为一个独立产品被 GCG 销售。正如其名所含的内容一样，SeqWeb 是威斯康星软件包基于网络的界面，通过如 Netscape Communicator 或 Internet Explore 等网络浏览器运行。基于网络的 GCG 界面在第 2 章中描述。

1.3.4 指南

本部分的指南说明 GCG 程序的基本用法，包括命名行及 SeqLab 两种方式的用法。由于所有的 GCG 程序操作方式相似，因此熟悉使用一种 GCG 程序则有助于使用另一种 GCG 程序。本指南一步一步地教你怎样用 Map 程序产生限制酶切图谱和从 GCG 核酸数据库索取的 DNA 序列中确定可读框。用 UNIX 操作系统和连接到 TCP/IP 网络的大多数服务器上进行少量改动就可直接用这些步骤工作。用户的登录环境应该设在 C 命令解释程序下。

1.3.4.1 所需的基本工具

为按照指南操作，除了 GCG 服务器的登录账号外，读者还需要有几个安装在个人计算机(Windows 或 MacOs)或 UNIX 工作站上的小工具以进行网络连接。所需的工具是远程终端程序(连接和控制 GCG 服务器用)和 X-Windows 服务器程序(在个人计算机屏幕上显示 X-Windows 图形和 SeqLab 窗口用)。其他的有用工具还有网页浏览器(在线阅读 GCG 手册用)、FTP 客户机程序(向/从 GCG 服务器传送文件用)、文本编辑器，如 Windows 的写字板程序(在个人计算机上编辑文本文件用)、打印工具(传送 PostScript 和 HPGL 文件到与个人计算机相连的打印机上)。作者使用的这些工具的地址信息见 1.4 节。GCG 网站上推荐的软件工具的其他信息也列于 1.4 节。

1.3.4.2 命令行操作指南

用个人计算机上的远程终端程序打开一个与 GCG 服务器的连接，并用预设的用户名(userid)和密码(password)登录，即会出现 UNIX 计算机命令提示符，如下所示：

```
UNIX%
```

在提示符后输入命令。用 UNIX 命令产生一个含 GCG 工作文件的文件夹，键入：

```
UNIX% mkdir gcg
```

并按回车键(仅键入 UNIX%后的字符)。键入下列字符，进入 GCG 目录。

```
UNIX% cd gcg
```

按回车键。键入下列字符，启动 GCG 程序。

```
UNIX% gcg
```

按回车键，GCG 程序欢迎标题滚动通过屏幕，显示可得到的数据库，片刻后，

回到命令提示符。

键入下列字符，用 GCG 取回(fetch)命令取回一份 m13mp18 克隆载体的 DNA 序列。

```
UNIX% fetch m13mp18
```

按回车键，文件名 m13mp18.gb_sy 进入 gcg 文件夹中。键入 UNIX 显示文件的命令证实该文件的存在。

```
UNIX% ls
```

按回车键。键入 UNIX 的 more 命令能检查 m13mp18.gb_sy 文件的内容。

```
UNIX% more m13mp18.gb_sy
```

按回车键。按空格键显示文件的下一页。文件扩展名 gb_sy 说明序列文件是来自 GenBank Synthetic 序列数据库。

对 m13mp18 序列运行 GCG Map 程序，键入：

```
UNIX% map m13mp18.gb_sy
```

按回车键。Map 程序开始运行，屏幕上出现程序的简短描述，然后提示用户怎样运行程序做出一些选择。对于本指南，并不每一步都选择，在提示符后简单地按回车键接受缺省选项。使得 Map 程序用整个序列(从碱基 1 到第 7249 位的最后一个碱基)搜索所有已知的限制酶切点(***)，以氨基酸单字母缩写翻译三个正向读框(*t*)，以 m13mp18.map 文件名保存结果到文本文件中。当程序运行时，点将滚动通过屏幕，程序运行完后，显示结果概要。UNIX 显示文件命令 ls 可以验证新结果文件 m13mp18.map 的存在。UNIX more 命令能如上文所述一样显示文件的内容。结果文件含有双链 DNA 序列，在 DNA 序列的顶上标有已知限制酶切点的位置。在序列的底部，有翻译读框中的单字母表示的氨基酸序列，其样式见图 1.4。如果需要，结果文件能打印(见 1.3.5 节)或用 FTP 传送到个人计算机上，输入到 Word Processor 程序或文本编辑程序中。

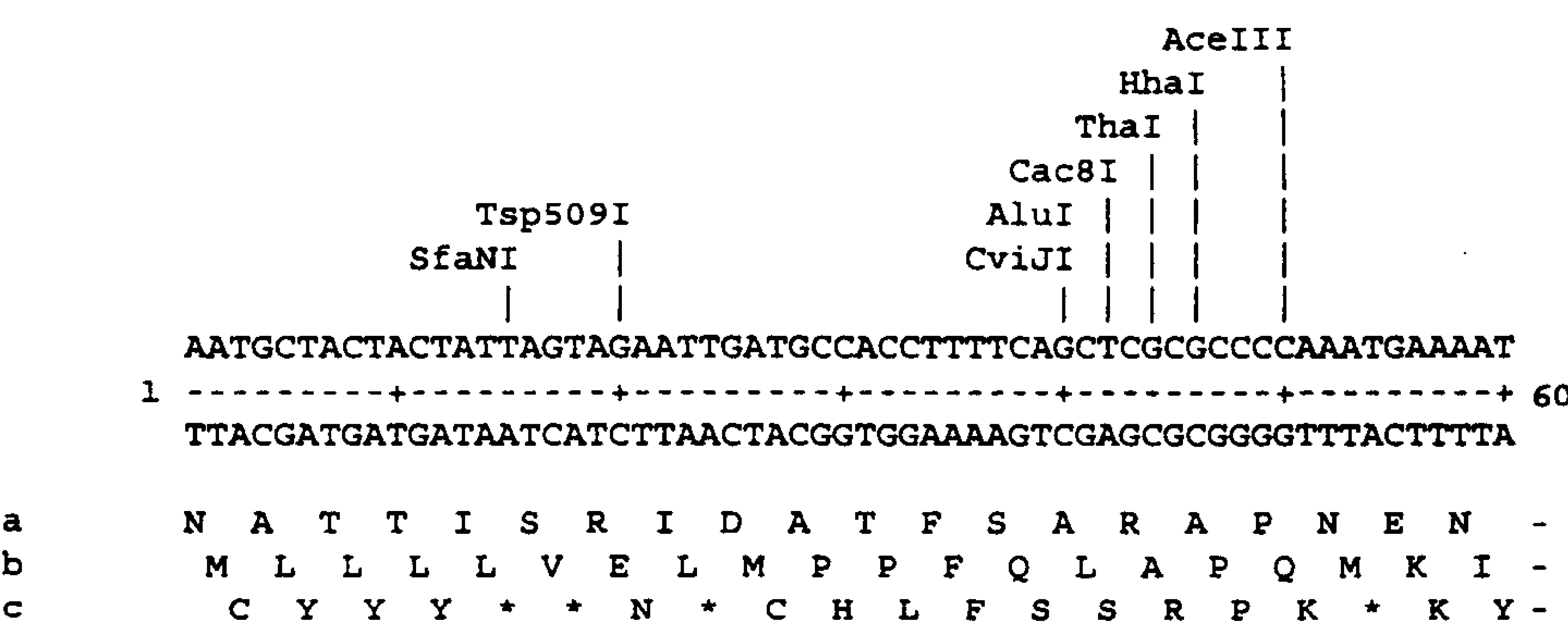


图 1.4 Map 程序输出结果

如果在命令行中，没有给定输入文件的名称就开始运行 Map 程序，程序的第一次提示就是输入序列文件名。输入的序列可以是用户的文件夹中的与本例一样的一个 GCG 数据库文件。在一个序列数据库中，直接运行 Map 程序，要在序列名前列上特定的数据库，如 gb_sy:m13mp18。这也正如本指南所采用的，在命令行中进行，也可以在程序开始后，回答输入序列的提示时进行(图 1.1)。在交互式程序对话中，用户可以在各种菜单选项中进行选择，控制程序操作方式。其他 GCG 程序的操作与此相似。

HYGCGmenu 程序也可用于产生 gcg 文件夹，从数据库取回 m13mp18 序列文件，运行 Map 程序，产生相似的结果。它不像上面所述的在命令行输入命令，而是用箭头指针键选择 Map 程序，把它指向输入序列文件并运行程序。在 HYGCG menu 内也能检查结果文件。

1.3.4.3 SeqLab 指南

SeqLab 指南中使用 X-Windows 模拟器。下列步骤不是设置 X-Windows 和 SeqLab 的唯一可用方法，而是在大多数情况下通用的一般方法。

在个人计算机上开始 X-Windows 服务器程序，并让它在后台运行。接着按照命令行指南的第一部分(见 1.3.4.2 节)所述方法，远程连接和登录 GCG 服务器，进入 GCG 文件夹，开始 GCG 程序。设置 X-Windows 的 DISPLAY(显示)环境需要个人计算机的 IP 地址(因特网)。如果仍不清楚，可在 GCG 服务器上得到，键入：

```
UNIX% who|grep userid
```

按回车键，命令行中 userid 替换成真实的用户名。用户登录的 IP 地址将显示在屏幕上登录 ID 行的右端。设置 DISPLAY 环境，键入：

```
setenv DISPLAY my.ip.address:0
```

按回车键，用真实的 IP 地址替换 my. ip. address，在最后不要忘记加上:0。这样 GCG 服务器上的 X-Windows 就显示在用户的个人计算机屏幕上，在这里以及本节接下来的第二段假设个人计算机上 X-Windows 程序已经启动。

启动 SeqLab，GCG 图形界面程序，键入：

```
UNIX% seqlab &
```

按回车键。“&”使程序在后台运行，以便需要时还可以输入其他命令。屏幕上弹出两个窗口，“SeqLab Main Window”(图 1.2)和“About SeqLab”窗口。在 About SeqLab 窗口中点击 OK 按钮，关闭该窗口。

SeqLab 对话中产生的文件保存在“工作目录”中。按下列步骤设定工作目录到 gcg 文件夹(在 1.3.4.2 节中命令行指南中生成)：在 SeqLab 主窗口中，点击 Options(选项)、点击 Preferences(参数选择)。在出现的新用户参数选择窗口中，点击 Working Dir...(工作目录)。双击 gcg folder(gcg 文件夹)，然后点击 OK，点击 Apply(应用)，然后再点击 OK。

SeqLab 通过使用一个序列列表文件工作。不同的项目可创建不同的列表文件。这里，本指南要创建的列表文件含一条序列，即 m13mp18 克隆载体序列。在 SeqLab 主窗口，点击 File(文件)、点击 New List(新列表)，输入 tutorial. list 并点击 OK。点击 File(文件)、单击 Add Sequences From(从……加序列)、点击 Databases(数据库)。在数据库 Specification(特性)中，输入 m13mp18、点击 Show Matching Entries(显示符合的数据条目)。在数据条目中，点击 m13mp18 进行选择，然后点击 Add to Main Window(加至主窗口)，点击 Close(关闭)。现在在列表中应该有一条数据 gb_sy:m13mp18。点击 File(文件)、点击 Save List(保存列表)保存形成的 tutorial. list 文件。工作列表可以含有来自数据库的序列条目，本例就是这种情况，也可以来自保存在用户工作目录或其他文件夹中的序列文件。

本指南的下一步是对 m13mp18 DNA 序列运行 Map 程序。如果未选序列，在 SeqLab 主窗口，点击列表中的 m13mp18 序列，使其变亮。依次点击 Functions、Mapping、Map，为 Map 程序打开一个新窗口。在 Map 程序窗口中，有选择各种选项的按键。在本指南，只需点击 Run 按钮开始 Map 程序，用缺省参数运行程序。在 SeqLab 主窗口中，先后点击 Windows 和 Job Manager，打开 Job Manager 窗口。Map 工作全部完成后，在 Job Manager 窗口报告，点击 Open Output Mgr，输出管理窗口打开。在输出管理窗口，选择结果文件，点击 m13mp18_nn. map，然后点击 Display，Map 程序的结果显示在屏幕上。出现的结果与 1.3.4.2 节中命令行程序运行 Map 程序得到的结果文件相似，不同的是所有 6 个读框都翻译了。如果需要，结果文件可以打印(见 1.3.5 节)或用 FTP 传送到个人计算机上，输入到 Word Processor 程序或文本编辑程序中。另一些 GCG 程序的操作与 SeqLab 相似。各个窗口能通过点击 Close 按钮关闭。SeqLab 主窗口能通过点击 File，然后点击 Exit 关闭。

1.3.5 显示和打印 GCG 图形和其他文件

很多 GCG 程序有图形输出，如限制酶切图或 RNA 二级结构预测。根据作者的经验，打印 GCG 图形结果是要掌握的重要技巧。然而，因为有很多不同的打印设备、终端程序、运行 GCG 程序的个人计算机，因此本节不可能包括所有可能的组合。这里的讨论将用作者熟悉的例子说明一些能用于打印 GCG 图形结果及其他文件的重要技术。这些例子是通过远程登录到 GCG 服务器连接个人计算机(Windows 或 MacOS)，所用 GCG 安装在 TCP/IP 网络中含 UNIX 操作系统的共享计算机上。尽管使用网络打印机也很方便，但这里所述的打印机是与 PC 机直接相连的兼容 PostScript 语言的打印机。

从 GCG 程序输出图形有几种选择。现描述如下：显示在屏幕上；用与 PC 机终端相连的打印机直接打印；保存输出为一个文件，然后打印。大多数 GCG 图形程序能以 PostScript 格式，也能以 HPGL 格式输出数据。打印结果需要打印机

兼容 PostScript 输出或 HPGL 输出。如果要通过网络直接在个人计算机相连的打印机上打印 PostScript 或 HPGL 文件，有必要使用兼容以“透明模式(transparent mode)”打印的终端程序，以便文件能直接送至打印机，而不用在 PC 机上操作。注意，这些方法不仅可用于 GCG 形成的文件，也可用于打印其他任何文件。如果没有 PostScript 兼容打印机，也可使用免费软件程序 GhostScript 和 GhostView 在 PC 机上显示和打印 PostScript 文件(见 1.4 节)。本节的第一个例子是从命令行或 HYGCGmenu 打印，后面的例子是从 SeqLab 打印。

1.3.5.1 GCG SetPlot 程序

在运行含图形输出 GCG 程序前，先用 GCG 的 SetPlot 程序告诉 GCG 怎样显示、打印或保存图形输出。该程序有一个菜单选择。这些选择由依赖于 GCG 系统员对 GCG 程序的设置决定。典型的选择如下。

1) X-Windows

选择 X-Windows 将在屏幕的一个窗口显示图形。如果 X-Windows 服务器程序已安装在 PC 机上或使用含内嵌 X-Windows 的 UNIX 工作站，该选项的任务能很好地进行。为了能顺利进行，必须先在终端上启动 X-Windows 程序，然后通过设定 DISPLAY(显示)环境参数(见 1.3.4.3 节指南)告诉 GCG 程序在哪显示它的 X-Windows。

2) PSFile 和 HPFile

这些选项分别保存图形为 PostScript 或 HPGL 语言文件。之后，可在任何 PostScript 或 HPGL 打印机上打印。要打印，应选下载至 PC 机并传送至打印机。在下载和拷贝到打印机的过程中，记住 PostScript 文件是简单的 ASCII(美国信息交换标准码)文件，而 HPGL 文件则是二进制码文件。打印说明见 1.3.5.3 节。

3) LW

用本选项，图形直接送到与 PC 机相连的 PostScript 打印机。打印机可以为任何 PostScript 打印机，而不仅仅是一台 LaserWriter。确认设定以透明模式从终端到打印机，为了在一页上打印完整的图形，可能有必要在打印机控制面板设置页保护。

4) Laser

用本选项，图形直接送到与 PC 机相连的 HPGL 打印机。打印机可以为任何 HPGL 打印机，而不仅仅是一台 LaserJet。确认设定以透明模式从终端到打印机，为了在一页打印完整的图形，可能有必要在打印机控制面板设置页保护。

5) Tek

一些远程登录程序已嵌入 Tektronix 模拟器。用此选项，在屏幕 Tek 窗口内显示图形，与 X-Windows 的例子类似。然而，对此选项，不需要设定 DISPLAY(显示)变量。

1.3.5.2 测试图形设置

使用 SetPlot 做出图形输出选择后，用 showplot 程序可列出图形设置。图形设置可用 plottest 程序进行测试，该程序绘出 GCG 测试图形。

1.3.5.3 打印 GCG 输出文件和其他文件

大多数 GCG 结果输出文件是纯文本文件，能在任何打印机上打印，也可导入任何文字编辑软件中。一些程序存储 HPGL 格式或 PostScript 格式的图形输出文件。这里有解释打印这些文件的选项的一些例子。注意，这些方法也能用于打印任何文件，即使这些文件不是 GCG 形成的。

1) 在与 PC 机连接的打印机上从 GCG 服务器目录打印一个简单文本文件

用 GCG 的 listfile 命令在与 PC 机或 Mac 机相连的打印机上打印任何文本文件。例如：

```
UNIX% listfile-noheading filename.txt
```

`-noheading` 命令行选项将在第一页的顶部不打印题头(文件名、日期)。

2) 在 PC 机相连的 PostScript 打印机上直接打印 GCG 服务器目录的简单文本文件。

用 GCG 的 lprint 命令在与 PC 机或 Mac 机相连的 PostScript 打印机上打印任何文本文件。在打印前，确认设置终端为以透明模式打印。例如：

```
UNIX% lprint-noheading filename.txt
```

`-noheading` 命令行选项将在第一页的顶部不打印题头(文件名、日期)。

3) 在与 PC 机相连的 PostScript 打印机上直接打印 GCG 服务器目录的 PostScript 图形文件

用含`-noheading`选项的 GCG listfile 命令可将 PostScript 图形文件在与 PC 机或 Mac 机相连的 PostScript 打印机上打印。在打印前，确认设置终端为以透明模式打印，为了在一页上打印完整的图形，也应该在打印机的控制面板上设置页保护。例如：

```
UNIX% listfile-noheading graphics.ps
```

4) 下载后在 PC 机上打印 PostScript 图形文件

首先用 ftp 将文件下载到 PC 机上。用 ASCII 文本模式转换 PostScript 文件。然后在 PC 机上，到 DOS 提示符下，拷贝文件到适当打印机接口的打印机上。例如：

```
C:\copy graphics.ps lpt1
```

Windows 用户可能喜欢用拖放式免费软件 PrFile 工具(见 1.4 节)。在 Mac 机上，用 LaserWriter 字型工具传送文件到打印机上。

5) 下载后在 PC 机上打印 HPGL 图形文件

首先用 ftp 将文件下载到 PC 机上。用 BINARY(二进制)模式转换 HPGL 文件。然后在 PC 机上，到 DOS 提示符下，拷贝文件到适当打印机接口的打印机上。例如：

```
C:\copy/b plot.hp lpt1
```

确认用/b 开关传送二进制文件到打印机。Windows 用户可能喜欢用拖放式免费软件 PrFile 工具。

1.3.5.4 打印 GCG SeqLab 的输出文件和其他文件

当从 SeqLabX-Windows 界面运行 GCG 程序时，输出文件既可以是文本文件，如 BLAST 搜索的结果，也可能是图形文件，GCG 图形程序或 SeqLab 能将这些图形文件作为图形在屏幕的 X-Windows 上显示。两种输出都能从 SeqLab 打印。除了打印任务被 GUI(图形用户界面)控制外，打印文法与上述命令行打印相似。

从 GCG 打印的一个重要条件是应该用具有良好打印能力的远程登录程序，如共享软件 QVT/Net 终端程序(见 1.4 节)。应该从 QVT/Net 终端程序启动 SeqLab(启动方法见 1.3.4.4 节指南)，注意这些方法可以打印任何文件，不只限于 GCG 形成的文件。

1) 在 PC 机相连的打印机上从 SeqLab 目录打印简单文本文件

SeqLab Output Manager Window(输出管理窗口)中所列的任何文本文件都可打印。所打印的文本文件可以是任一 GCG 程序的输出文件，也可以是用 Add Text File...按钮加到输出管理窗口中的任何文本文件。要打印某文本文件，在列表中点击其文件名，使其加亮，点击 Print 按钮，在 Output Format 下选择 ASCII。在 ASCII Print Command 下选择或键入：

```
listfile-noheading
```

然后，点击 OK。将用 GCG listfile 命令打印该文本文件。加`-noheading`开关将在第一页的顶部不打印题头(文件名、日期)。

2) 保存 SeqLab 的图形为 PostScript 文件

SeqLab 输出管理窗口中的任何图形文件都能保存为 PostScript 的文件。如果图形文件未在输出管理窗口中列出，能用 Add Graphics File...按钮加到列表中。为了保存图形到一个文件，在列表中点击文件名使其加亮，然后点击 Print 按钮。在 Output Device 中，选择 PostScript output saved as file graphics. ps，在 Port of File 下选择或键入：

```
graphics.ps
```

或键入任何要保存的文件名。如果同时运行多次，用不同的文件名保存每个文件。接着，点击 Proceed。文件将保存在 GCG 服务器硬盘中 SeqLab 启动的文件夹中，文件能按上述 PostScript 文件方法[见 1.3.5.3 节的 3)]下载或打印。

3) 连到 PC 机的打印机上打印 SeqLab 的 PostScript 图形文件

先按上述的方法保存图形(即图形文件)为 PostScript 文件。然后点击 Add Text File...按钮(记住 PostScript 文件仅是文本文件，PostScript 打印机能解释并打印为图形)，加 graphics. ps 文件到你的输出管理窗口。然后按下述操作：设置远端程序为透明模式，选 PostScript 兼容打印机，文本文件用 `listfile-noheading` 命令，再按上述方法[见 1.3.5.4 节的 1)]打印。

4) 保存 SeqLab 图形为 HPGL 文件

任何位于 SeqLab 输出管理窗口中的图形文件都能保存为 HPGL 文件。如果图形文件不在输出管理窗口的列表内，可用 Add Graphics File...按钮加到列表中。为了保存图形为文件，在列表中点击文件名使其加亮，然后击 Print 按钮。在 Output Device 下，选择 HPGL output saved as file plot. hp，再在 Port or File 下选择或输入：

```
plot.hp
```

或输入任何其他要保存的文件名。如果同时运行多次，用不同的文件名保存每个文件。接着，点击 Proceed，文件将保存在 GCG 服务器硬盘 SeqLab 启动的文件夹中，文件能按上述 HPGL 文件方法[见 1.3.5.3 节的 4)]下载或打印。应强调的是，HPGL 文件是二进制文件。

1.4 注

(1) 关于威斯康星软件包的更多信息可在遗传学计算机小组的网站得到：`http://www.gcg.com/`。

(2) GCG HYGCGmenu 超文本菜单及伴侣程序 HYBROW 可以从下列网址得到：`ftp:biomed.nus.sg/pub/biocomp/`。

(3) 作者使用 QVT/Net Terminal(远程登录)程序是因为其透明模式打印(机器转移归向打印)GCG 图形的能力。它也能从命令行或 SeqLab(如果 SeqLab 从 QVT/Net Terminal 对话中启动 SeqLab)内打印好 GCG 图形。关于共享软件 QVT/Net Terminal 程序的进一步信息可见于：`http://www.frontiernet.net/~qpcsoft/`。

(4) 免费终端程序 TeraTerm 能在屏幕上显示 Tektronix 图像。质量不如用 X-Windows 或打印 PostScript 或 HPGL 文件高，但足以看出图意。最新版的 TeraTerm 可从下列网址得到：`http://hp.vector.co.jp/authors/VA002416/teraterm.html`。

(5) 作者用拖放式 PrFile 打印工具传送 PostScript、HPGL、二进制(如*.prn 文件)和文本文件到打印机(从服务器下载到 PC 机后)。Windows 用免费软件 PrFile 打印工具可从下列网址得到：`http://hem1.passagen.se/ptlerup/prfile.html`。

(6) 没有 PostScript 兼容打印机时，免费软件 Ghostscript、Ghostview 及 Gsview 能用于显示和打印 PostScript 文件。Ghostscript、Ghostview 及 Gsview 可从下列网址下载：`http://www.cs.wisc.edu/~ghost/`。

(7) Micro X-Win32 X 服务器程序能运行 GCG 的 SeqLab 图形用户界面及其

他 UNIX X 程序，如 File Manager(文件管理器)或 ClustalX。关于 Micro X-Win32 X-Windows 服务器程序更多的信息及其免费演示版可从下列网址得到：http://www.starnet.com/docs/xwin32.html。

(8) 最新版共享软件 WS_FTP FTP 客户端软件可从下列网址得到：http://www.ipswitch.com/。

(9) 最新版 Netscape communicator 网页浏览器软件可从下列网址得到：http://home.netscape.com/computing/download/index.html。

(10) 最新版 Internet Explorer 网页浏览器软件可从下列网址得到：http://www.microsoft.com/ie/。

(11) UseNet GCG 讨论组 INFO-GCG/bionet. software. gcg 位于：http://www.bio.net/hypermail/INFO-GCG/。

(12) 作者的主页可在下列网址找到：http://cmmg.biosci.wayne.edu/。

(13) 作者与上列任何软件的提供者没有从属关系。

(欧阳红生　译)

参 考 文 献

[1] Devereux, J., Haeberli, P., and Smithies, O. (1984) A comprehensive set of sequence analysis programs for the VAX. *Nucleic Acids Res.* **12,** 387-395.

[2] *Wisconsin Package Version 9.1.* Genetics Computer Group (GCG), Madison, WI.

[3] Zuker, M. (1989) Computer prediction of RNA structure, in *Methods in Enzymology*, vol. 18. (J.E. Dahlberg and J.N. Abelson, eds.), Academic, San Diego, CA, pp.262-288.

2 GCG 序列分析程序基于网页的界面

David D. Womble

2.1 引言

遗传学计算机小组(GCG)程序，又称为威斯康星软件包，含功能强大的操作、分析和比较核酸与蛋白质序列的整套软件工具。威斯康星软件包含有 130 多个程序，每个都是完成特定任务的工具，如翻译核苷酸编码序列或确定限制酶切位点。威斯康星软件包一般安装在网络的共享计算机上。传统上，用户用一个终端程序登录和操作 GCG 程序。近来的努力导致建立了威斯康星软件包基于网页的界面，它使得 GCG 程序能从网页浏览器，如 Netscape Communicator[见 2.4 注(6)]或 Internet Explore[见 2.4 注(7)]运行。由于网页浏览器现在很普遍且特别易用，因此该界面为大多数人使用 GCG 程序提供了一种熟悉和轻松的操作方式。GCG 近来推出的基于网页的 GCG 界面称为 SeqWeb。SeqWeb 作为已安装的 GCG 的附件与威斯康星软件包分开提供[见 2.4 注(3)]。另一种基于网页的界面名为 BioPortal[见 2.4 注(4)]，由新加坡国立大学开发。BioPortal 也作为已安装的 GCG 的附件。其他的基于网页的 GCG 界面也在研究中，这里就不提了。

本章的介绍假定读者已经熟悉威斯康星软件包。关于 GCG 程序本身的讨论见第 1 章。

2.2 材料

这里报道的方法是 GCG 程序软件包 9.1 版，安装在通过 TCP/IP 网络连接的含 UNIX 操作系统的共享计算机上。软件包可安装在几种不同的计算机上，包括运行 Digital UNIX 4.0 的 Digital Alpha 机器，运行 6.2、6.3 或 6.4 版 IRIX 的基于 RISC 的 Silicon Graphics 机器，及运行 2.51 或 2.6 版 Solaris 的基于 SPARC 的 Sun 机器上。安装和维持含整套数据库的威斯康星软件包需要最少 15G 字节的硬盘空间。还需要存放个人用户文件的额外的硬盘空间。应该有最少 128M 的核心内存和 200M 的虚拟内存。

SeqWeb 是操作 GCG 程序的基于网页的界面。SeqWeb 产品包括网页服务器

软件。它仅运行在上面所列的基于 UNIX 的计算机上。SeqWeb 必须安装在已安装并运行了威斯康星软件包的服务器上。这里介绍的是 SeqWeb 发布前的版本。最终的产品后来已经发布。安装 SeqWeb 需要大约 16M 硬盘空间。威斯康星软件包和 SeqWeb 可以从遗传计算机小组得到，其地址为：Genetics Computer Group, 575 Science Drive, Madison, WI53711; (608)231-5200, 传真: (608)231-5202; 电子邮件: info@gcg.com, 网址: http://www.gcg.com。

由新加坡国立大学开发的 BioPortal 是另一个操作 GCG 程序的基于 Web 界面。与 SeqWeb 一样，它含网页服务器软件，必须安装在 GCG 已安装并运行的 UNIX 服务器上。安装 BioPortal 需要大约 17M 的硬盘空间。BioPortal 的更多的信息可从 BioPortal 网页站点得到。其站点地址为: http://bic.nus.edu.sg:8888/, 或发电子邮件到: meena@bic.nus.edu.sg。

为了操作基于网页的 GCG 界面，需要一台通过 TCP/IP 网络相连的个人计算机(Windows 或 MacOS)或 UNIX 工作站。这些计算机或工作站应安装 Netscape Communicator 4.0 或更高版本，或 Internet Explorer 4.0 或更高版本。基于网页的 GCG 界面使用 Java 语言，这是需要最新版本浏览器的原因。对每个 GCG 服务器的用户，必须设置含用户名和密码的访问账户。

2.3 方法

2.3.1 GCG SeqWeb 界面

2.3.1.1 SeqWeb 概述

SeqWeb 是威斯康星软件包序列分析工具的附件，它为 GCG 程序提供了基于网页的界面。使用 Netscape 或 Internet Explorer，简单地登录 GCG 服务器就可访问威斯康星软件包。从菜单中选择 GCG 程序，用复选框、下拉菜单、文本框等确定运行程序的选项。

有几种方法将序列装载到 SeqWeb Work Area(SeqWeb 工作区)。可以浏览 PC 硬盘上的文件，将序列从用户的 PC 机上传到服务器；也可从剪贴板粘贴数据；还可在服务器的 GCG 数据库内选择序列。序列文件一旦载入 SeqWeb，就存放在服务器硬盘用户分得的空间内。SeqWeb 自动翻译其他格式的文件，如 FASTA、EMBL 及 GenBank 格式文件供 GCG 程序操作。进入 SeqWeb Sequence Manager(SeqWeb 序列管理)功能允许对序列文件进行增加、编辑、改名、拷贝或删除。

运行 GCG 程序的结果也储存在服务器硬盘用户分得的空间内的文件中。文本和图形结果用彩色显示在屏幕上，也能打印或保存到用户 PC 机的硬盘上。文本输出可以以 HTML 格式保存，这对结果含内部数据库及网页上的外部数据连接特别有用。当程序结果列出了数据库序列时，选择感兴趣的序列会使 SeqWeb 将

它们装载进 Work Area(工作区)。结果文件能用 SeqWeb Results Manager 管理，它含有对结果文件显示、编辑和删除的功能。它易于进入和使用，加上 GCG 程序的强大功能，形成了一个非常有效的全面的软件包。

2.3.1.2 SeqWeb 外观

GCG 界面 SeqWeb 使用框架，浏览器的左边有 Content(内容)框架，右边为 Work Area(工作区)框架，底部为 Utilitiy(工具)框架(图 2.1)。内容框架能被锁定显示基于菜单的程序功能，或所有可得到程序的索引。点击任一程序的功能，如 Mapping 或一个程序名如 Map，在新工作区框架内打开新内容。工作区框架内含

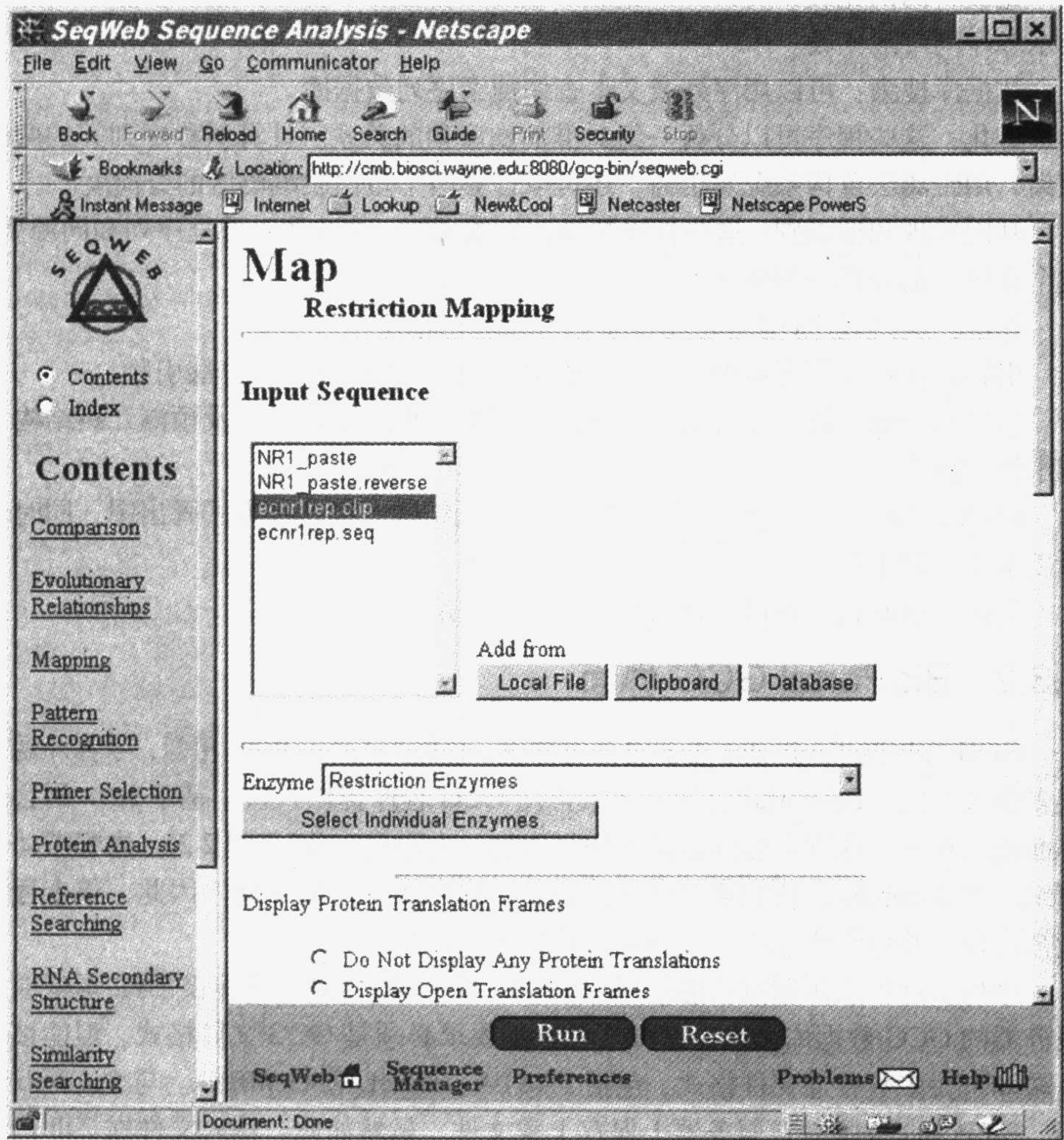


图 2.1 GCG 程序 SeqWeb 界面

有的项目，如按钮、复选框、下拉菜单、文本框、序列文件列表等，它们被用于操作给定的 GCG 程序及选择要运行的序列。底部的工具框架内容根据工作区框架内容而改变。工具包括进入序列管理、项目管理、结果管理、用户优先选择设定及一个 Help 按钮，用于在线打开 GCG 文件。当程序载入工作区框架内时，按钮 Run 和 Reset 也位于工具框架。

2.3.1.3 SeqWeb 内可得到的 GCG 程序

SeqWeb 提供了威斯康星软件包中最常用的一些程序的入口，这些程序属于下列功能类中。

比较：两两比较，可以对准比较，也可以点图(dot-plot)比较；多重序列比较显示多序列对准比较。

数据库搜索：可以用序列及文本查询搜索序列数据库。

进化：这些程序可以分析一组对准前的序列的关系。计算对称序列间的两两距离，用距离法重构系统发生树，还可以计算两个蛋白质编码区的分离度。

基因发现和型识别：这些程序帮助识别编码区、终止子、重复序列和同源型，数个程序帮助分析序列成分。

作图：这些程序计算和显示限制图和肽切割图。

引物选择：此程序帮助从一个模板 DNA 序列选择寡聚核苷酸引物。

蛋白质分析：该程序分析特异蛋白质序列，识别序列基序及预测二级结构、疏水性和抗原性。

RNA 二级结构：这些程序预测 RNA 二级结构，以几种方式画出图，同时也可以鉴定反向重复。

翻译：这些程序可以翻译核酸为蛋白质，反之亦然。

2.3.2 BioPortal GCG 界面

GCG 程序的 BioPortal 基于网页的界面，总体上与 SeqWeb 相似，尽管外观和很多细节不同。BioPortal 也为大多数常用 GCG 程序提供入口。基于 Java 语言的界面使用框架，在浏览器屏幕的左边是内容，右边是工作区(图 2.2)。内容按功能组织，如 Mapping，功能菜单可以扩充显示程序排列，如 Map。因此，单个程序能被选择，并可将所选程序装载到右边的工作区。

序列文件可以用浏览功能从用户的 PC 机上载，从剪贴板粘贴到文本框或从服务器的 GCG 数据库调出。序列文件并不要求必须是 GCG 文件格式，但能被迅速转译自大多数普通文件格式，包括 FASTA、EMBL 和 GenBank。与 SeqWeb 不同，BioPortal 仅临时在服务器上保存上载文件。虽然如此，输入文件在 24h 内可由另一 GCG 程序使用。

一旦一个序列被载入程序，程序的选项就能用复选框、单选按钮、文本框等

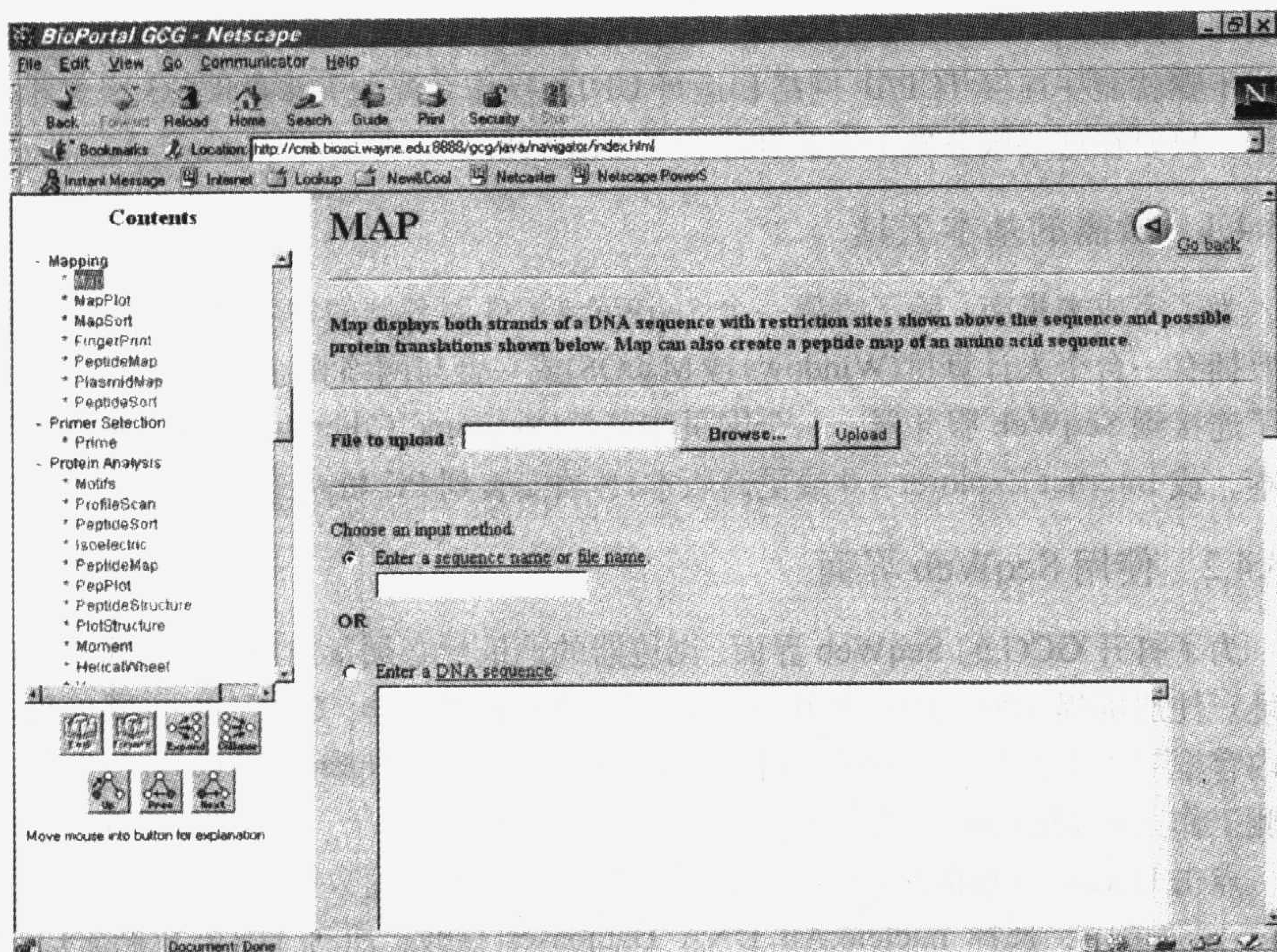


图 2.2　GCG 程序的 BioPortal 界面

选择，然后，程序可通过按 Submit 按钮启动，随后，程序的结果出现在屏幕上。结果能被显示、打印或保存到用户的 PC 机硬盘上。结果文件也被临时存放在服务器上，以便在适当的时候能提交给其他 GCG 程序进行更多的分析。

除了作为 GCG 程序的界面外，BioPortal 也兼容一系列序列分析程序，包括 CLUSTALW、PHYLIP、Primer 和 ReadSeq 等程序。那些程序的每一个也有一个基于网页的界面包括在 BioPortal 内。与 GCG 程序的界面一起，在一个便利的场所提供了一套重要的序列分析工具，用易用的网页进入所有这些程序。

2.3.3　GCG 其他的基于网页界面

尽管本章没有涉及，GCG 程序还有其他的基于网页界面。两个例子是由比利时的 EMBNet Node 开发的 WWW2GCG 及英国 Hinxton 欧洲生物信息学研究所(EMBL-EBI)和德国海德尔堡德国癌症研究中心(DKFZ)合作开发的 W2H。关于怎样得到更多的这些界面的信息见 2.4 节注中。

2.3.4　SeqWeb 指南

这部分的指南解释通过网页界面使用 GCG 程序的基本用法。本指南只介绍 SeqWeb，但与运行 BioPortal 过程很相似。逐步地讲解说明怎样用 Map 程序产生

限制图和测定从 GCG 核酸数据库取回的 DNA 序列可读框位置。只要很少的改变，这些步骤就能在连接 TCP/IP 网络和运行 UNIX 操作系统的大多数 GCG 服务器上工作，只要在这些服务器上安装和运行了 SeqWeb。

2.3.4.1 所需的基本工具

为了完成本指南，除了需要一个 SeqWeb/GCG 服务器的登录账号外，读者还需要拥有一台个人计算机(Windows 或 MacOS)或一台与网络相连的 UNIX 工作站。为了连接到 SeqWeb 服务器，一个网页浏览器(Netscape Communicator 4.0 或更高版本，或 Internet Explorer 4.0 或更高版本)必须安装到 PC 机或工作站上。

2.3.4.2 使用 SeqWeb 界面

为了打开 GCG 的 SeqWeb 界面，浏览器的地址栏必须指向 SeqWeb 服务器的地址。注册屏提示输入用户名和密码。一旦登录上服务器，SeqWeb 主屏出现。在内容框架中点击 Index 按钮直到 Map 出现，然后点击 Nucleic(核酸)按钮，打开右侧工作区的 Map 程序(图 2.1)。

点击 Database(数据库)按钮，它打开从数据库加入序列数据屏幕。在 Search(搜索)下拉菜单中，选择 nucleic:All DNA Databases(核酸：所有 DNA 数据库)。在 Sequence locus name or accession number(序列基因名称或登录号文本框中，输入 m13mp18 然后点击 OK。在回来的屏幕上，点击记录旁边的复选框选择 gb_sy:m13mp18 记录，然后点击 Add Selected。点击 Close 关闭此窗口。在 Map 程序框架内，gb_sy:m13mp18 序列已加到 Input Sequence(输入序列)列表中。

如果仍未加亮，在列表中点击 gb_sy:m13mp18 序列，以选中序列。各种可选的运行参数可以修改。但是在本指南中，仅点击 Run 按钮用全部缺省参数提交任务，结果将出现在屏幕上，有两条 DNA 序列，已知限制酶切点位置标在 DNA 序列的上面，单字母表示的翻译的可读框的氨基酸序列位于 DNA 序列的下面(图 2.3)。

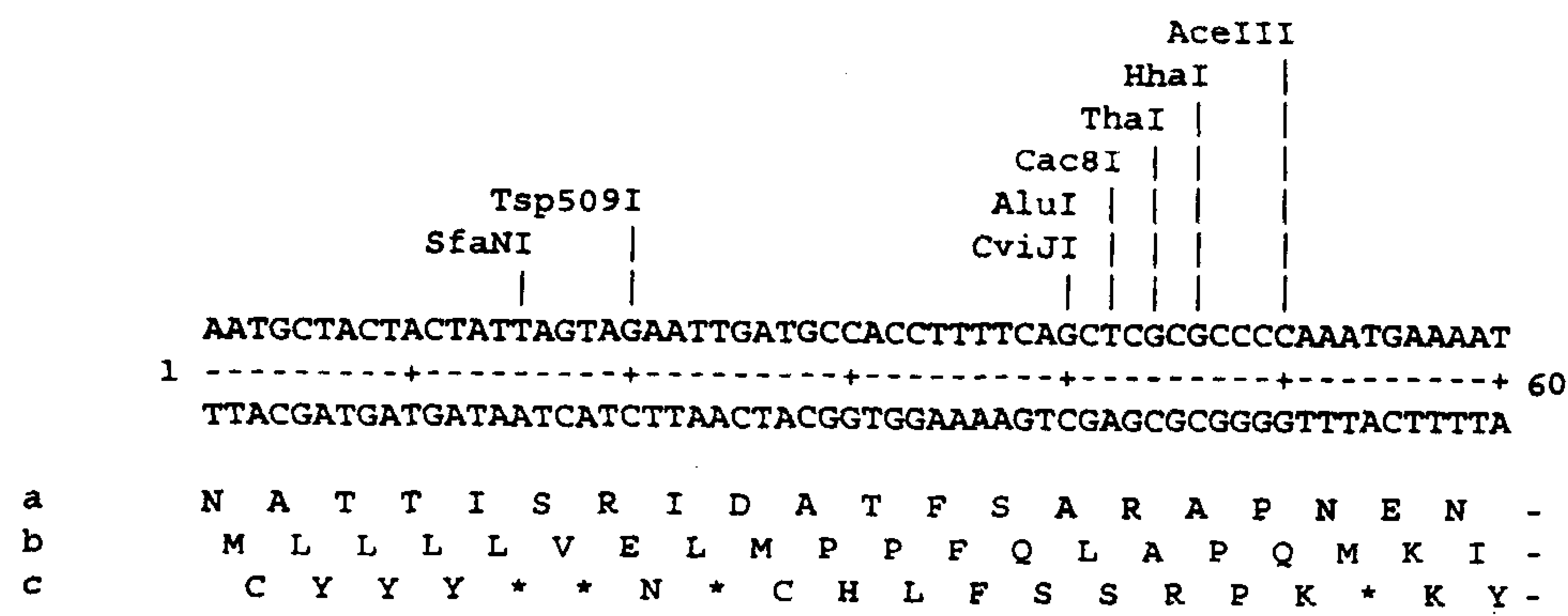


图 2.3 Map 的输出结果

结果可以打印或者以 HTML 或纯文本格式保存在用户的 PC 机硬盘上；结果也可以类似于 m13mp18_map_14885.htm 的文件名的文件保存在服务器的硬盘上，且以后能在 SeqWeb Results Manager 中索回。输入序列文件 gb_sy:m13mp18 能通过点击相连的结果屏的顶部显示，或者点击底部附近的结果屏，滚动显示序列框。其他 GCG 程序操作与此类似。

2.4 注

(1) 关于威斯康星软件包和 SeqWeb 更多的信息可从遗传学计算机小组的网站上得到。网站地址为：http://www.gcg.com/。

(2) 关于 BioPortal 更多的信息可从下列网址得到：http://bic.nus.edu.sg:8888/。

(3) GCG 基于网页的界面 W2H 的信息可从下列网址得到：http://industry.ebi.ac.uk/w2h/。

(4) GCG 基于网页的界面 WWW2GCG 的信息可从下列网址得到：ftp://alize.ulb.ac.be/pub/www2gcg/。

(5) GCG 的 UseNet 讨论组 INFO-GCG/bionet. software. gcg 位于：http://www.bio.net/hypermail/INFO-GCG。

(6) 最新版 Netscape Communicator 网络浏览器软件可从下列网址得到：http://home.netscape.com/computing/download/index.html。

(7) 最新版 Internet Explorer 网络浏览器软件可从下列网址得到：http://www.microsoft.com/ie/。

(8) 作者的主页可在下列网址找到：http://cmmg.biosci.wayne.edu/。

(9) 作者与上列任何软件的提供者没有从属关系。

(欧阳红生　译)

3 Omiga：一种基于 PC 机的序列分析工具

Jeffrey A. Kramer

3.1 引言

基于计算机的序列分析、注释和操作是所有分子生物学工作者对大多数简单的 DNA 序列处理所必需的。由于序列数据的不断增加，对单个序列以及序列元件进行操作和注释的工具对分子生物学工作者来说更显得至关重要。1.1 版本 Omiga DNA 和蛋白质序列分析软件，提供了一个有效且全面的核苷酸和蛋白质序列分析工具，该工具普遍适用于标准的 PC 机。Omiga 可以从几种常见的格式文件中输入序列。序列输入并将其分配给不同的项目后，Omiga 允许使用者创建、分析和编辑序列比对结果。并且可对序列进行限制性位点、基序及其他序列特性搜索，这些结果都被添加到每一个序列的注释中。最后，Omiga 还提供快速查找推导的编码区以及 PCR 和测序引物。

本章的重点是介绍 Omiga 提供的主要功能，以及输出的类型和方法。仅用一章的篇幅来详细阐述如何操作 Omiga 是很困难的。事实上，Omiga 还有一个操作手册，350 多页[1]。这个手册十分有用，它详细讲解了如何进行操作。它简单易学，包括一些逐步学习指南，这些指南对于学习 Omiga 界面和熟悉操作十分有用。

3.2 材料(硬件要求)

1.1 版本的 Omiga 设计运行于 Windows 95 或 Windows NT 上，需要 CD-ROM 光驱和 17M 的硬盘空间。至少还需要大约 15M 的空间安装外围工具，包括 VecBank 序列文件、Rasmol 和 Adobe Acrobat。还提供详尽的使用指南及一些在线指南，这需要 1M 的硬盘空间，该软件运行于 233MHz AMD K6 处理器上，机器配置稍差也可运行。

3.3 普通程序界面

3.3.1 项目

Omiga 采用项目管理概念(project concept)组织序列及由程序分析得出的各个

序列的注释、附加结果和信息。一个项目代表一种组织研究者资料的简单方法，可将不同的项目安装到硬盘的不同区，甚至一个远程存储装置，如 Zip 驱动器。多位用户也能够用单独的项目来组织无关序列和来自实验室中无关联项目的序列，以减少每一个界面中出现的多余资料，这样，零星使用的用户在大多数情况下无需打开多个项目。

3.3.2 项目浏览

3.3.2.1 概述

使用 Omiga 的第一步包括建立一个项目，在打开的 Omiga 上方的工具栏中选择 New Project(新项目)按钮即可。在计算机里这个 New Project 文件夹已经分配了一个名字和一个文件。一旦这个新的项目被定义，这个项目浏览就被启动。在 Omiga 中，Project View(项目浏览)是一个主窗口(图 3.1)，该窗口提供了该项目所有序列和相关数据的总览。在这个窗口中，用户能够按照要求创建另外的文件夹和子目录来管理数据和信息。例如，可按照要求将肽链和核苷酸序列分别存入不同的文件夹，以及把经 Omiga 处理得到的比对结果存入另一个文件夹。

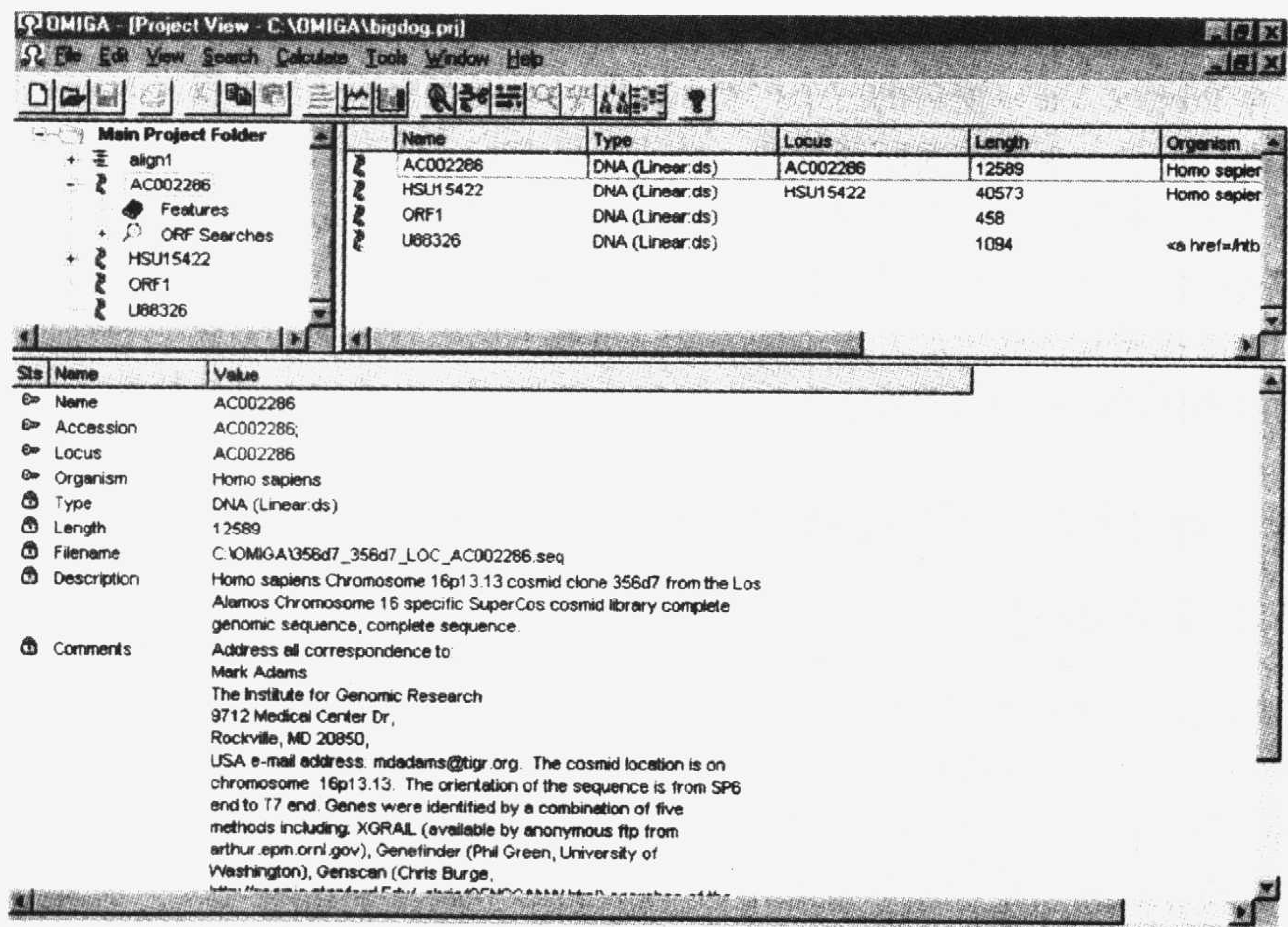


图 3.1 项目浏览窗口提供所有项目中序列和其他数据的概观

3.3.2.2 树式控制面板(tree panel)

用户可通过 Project View 窗口快速查看大量的数据。它包括 3 个子窗口，与 3

个控制面板相联系。在 Omiga Project View 窗口的左边是树式控制面板，此窗口显示一个项目中单个对象的各级目录情况。点击文件夹左边的“+”标志，能够打开文件夹，显示该文件夹包含的内容。每一个对象的类型由它的名字以及所具有的特征性的图符指明。例如，主项目以及下一级目录都由文件夹的图符指明，而 DNA 和 RNA 序列则分别显示为双链或单链。这个图符甚至能够表明 DNA 是环型或是线型，这点对限制性位点搜索有用。该图标能使用户快速查看整个项目中每一个对象所包含的内容。

3.3.2.3 总览和属性面板(summary and attribute panel)

在树式面板中打开文件夹，文件夹所包含的项目就展示在总览面板中。总览面板位于 Omiga Project View 窗口的右边。在树式面板中展开的文件夹项目，能够在总览面板中显示文件夹中该项目的更详细列表。例如，展开包含 DNA 序列的文件夹，将会在总览面板中显示序列的名称、大小和登记号(如果有的话)，及一些附加信息，如序列的来源。选中总览面板中的一个对象(可以通过点击树式面板中的一个对象来做到)，在第三个控制面板中显示该对象的属性。Project View 窗口中的第三个控制面板是属性面板，它位于项目浏览的下方，可显示被选中对象的所有属性。这个视图看起来很像 GenBank 序列视图。

在 Project View 窗口的三个控制面板能够通过点击和拖拉面板之间的边框来调整窗口大小。正如前面所谈到的，树式面板中的文件夹通过点击文件夹图符左边的“+”来浏览文件夹的内容。同样，单个序列的附加信息也可照此浏览。和文件夹图标一样，具有附加信息的序列在其图标的左侧也具有“+”符号。点击它就可以在树式面板中展开更多的分支，可显示该项目附加的数据，包括用户生成的限制性图谱和功能图谱，在后面将要介绍如何产生和浏览这些特性。

3.4 项目管理和程序组织

3.4.1 序列浏览

项目以及项目内所有的元件的管理很容易，可以通过不同的浏览窗口进行。正如在上面 3.3 节中所介绍的，Project View 窗口是 Omiga 的主要浏览窗口。然而，一些其他的浏览窗口也可以打开，这些窗口都可由 Project View 窗口直接进入。双击 Project View 窗口的 Tree View 中的单个序列，能够打开这个序列的 Sequence View(序列浏览)。利用 Sequence View 可手动编辑序列，这个特性可用来记录实验室序列引入的特异突变位点和序列多态性。经过编辑的序列可用来鉴别新的限制性位点。编辑功能还用于将一个序列的大片段插入到另一个序列中，反映实验室克隆实验的情况。因此，使用 Omiga 程序，可以模拟将数据库中的序列插入到质粒载体中的克隆实验，质粒序列和插入片段序列是已知的。这个新的序列可用来

查找限制性图谱，该图谱可验证实验室中实验的操作是否正确。这个新构建的序列和所有的分析能够以一个新序列的形式存入到这个项目中。另外，与构建新序列所用的原始序列片段相关的所有的特性(如复制原点和氨苄青霉素抗性基因)将成为新序列的一部分。用户可通过 Sequence View 对序列进行翻译和反向翻译，产生互补链及反向互补链。

3.4.2 比对浏览

Omiga 的另外一个可利用的浏览窗口是 Alignment View(比对浏览，见 3.6 节)。正如其名称所表明的，对多个序列进行操作和浏览序列比对时可利用比对浏览窗口。当进行比对时，可从 Project View 中的树式面板中选择多个序列。Align Sequences(比对序列)选项可从工具栏中的 Calculate(计算)选项的下拉菜单中选取。3.6 节提供关于序列比对的附加信息。比对浏览窗口包含总览面板，总览面板能够显示序列的名称和显示面板(display panel)，显示实际的比对结果，能够进行多达 500 个长度在 10 000 以内序列的比对。过多的序列比较可引起 CPU 的过度运算导致操作系统不稳定以致死机。比对浏览也可用于进行序列组合(group sequence)、添加或编辑缺口位置(gap position)，以及为比对添加注释。最后比对一致的结果作为该单个项目数据存储。

3.4.3 查找结果浏览

利用 Search Results View(查找结果浏览)窗口，可以以表格的形式显示 Omiga 所得到的多种查询结果(见 3.7 和 3.8 节)。执行一次查询可以产生一个查找结果浏览窗口(见 3.7 和 3.8 节在 Omiga 中执行查找的附加信息)。查找结果浏览窗口可以按照许多标准以表格的形式显示结果，并符合单个用户的要求。因此限制性位点按照降序的顺序，按频率从高到低显示。也可按照要求显示序列中特定位置限制性位点，用户可利用别的命令显示其他查找结果。查找结果可以进行筛选以减少非必需的信息。如某些限制性位点可被筛选掉，只保留感兴趣的位点。在查找结果浏览窗口中显示的数据能够在该项目中存储或被直接打印出来。

3.4.4 特性浏览及项目浏览窗口

Features View(特性浏览)窗口是最有用的浏览窗口。特性浏览窗口以图的形式显示应用 Omiga 得到的不同的查找结果(见 3.7 和 3.8 节)。以可视化的图形代表序列，在图上，序列的不同限制性位点和特性被标记在上面。例如，在特性浏览窗口中以水平的线表示约 53kb 人精蛋白基因(accession nos. U15422 和 AC002286)。然而，由于各种特性类型的出现，屏幕上表示的信息日益增多(图 3.2)。对基因可进行多重表示，如可用单箭头表示整个基因，或将基因断裂成外显子进行表示。这种多重表示还可以加注一些信息，如编码区限制性片段，以及其他重要的序列

元件。尽管应用 Omiga 识别重复序列元件比较困难，但可进行手动添加项目中的序列的重复元件，当然被添加的位置是已知的。其他的分析工具，如 Censor server(`http://charon.lpi.org/~server/`)能够识别重复序列元件，并能够提供它们在序列中的准确位置。另外，如果序列来源于某一个数据库，且这些元件已经存在于序列中，可将这些元件输入到 Omiga 中。就像 Project View(项目总览)窗口一样，特性浏览窗口是由多控制面板组成的，包括树式面板和显示面板。树式面板用来切换具有不同的重叠组的特性的开和关，而显示面板包含在树式面板中处于打开状态元件表示的具体信息。

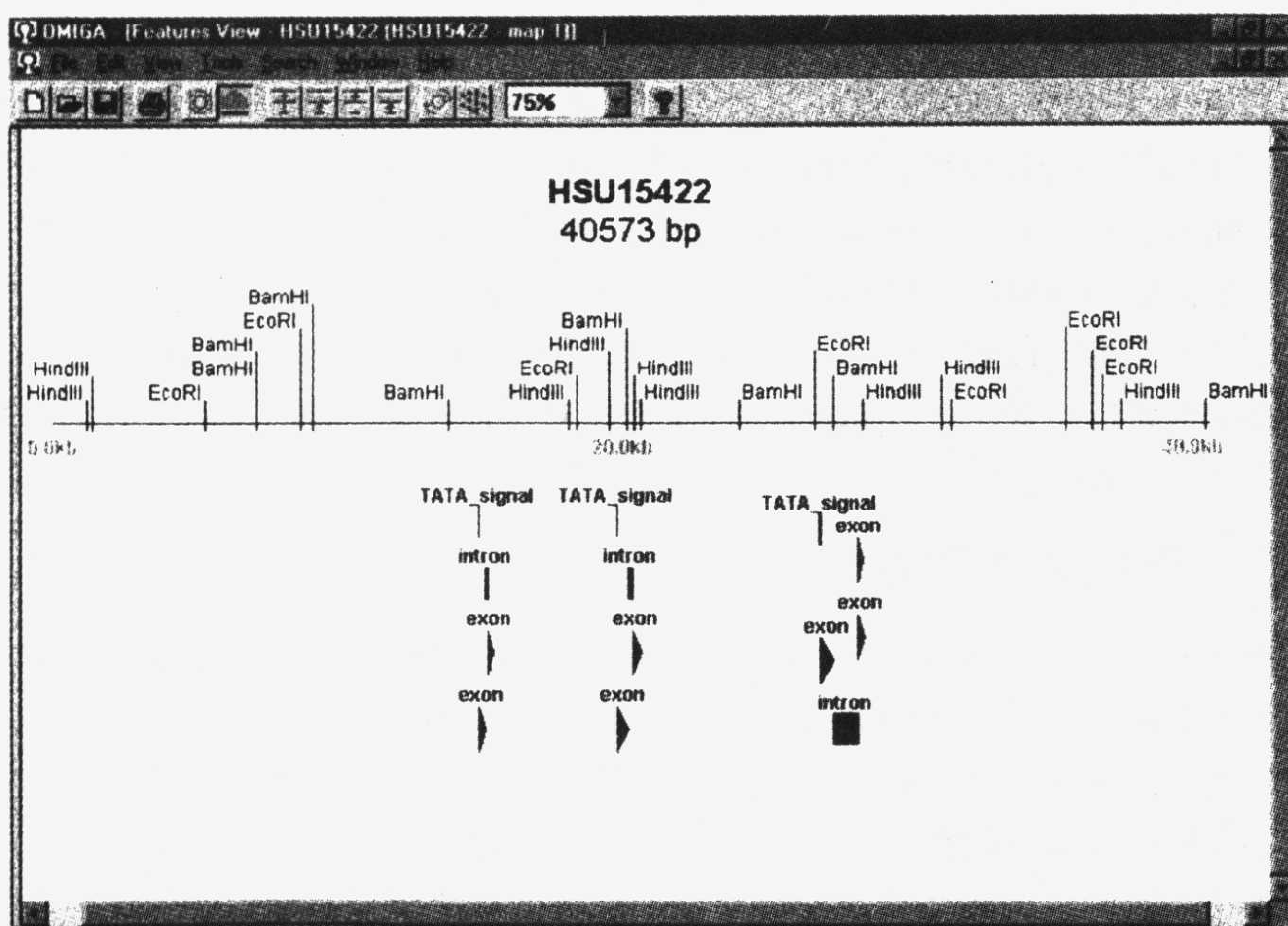


图 3.2 Features View 提供序列的一种可视化的表述

3.4.5 特征与组成分析浏览

在 Omiga 中最后的两个浏览窗口是 Profile View(特征浏览)和 Composition Analysis(成分分析)窗口。当从 Omiga 窗口上方的工具栏中选定 Calculate 选项时，可从下拉菜单中分别选取 Property Profile(特征)或 Composition Analysis(成分分析)来显示一个或多个序列的特征或成分分析浏览窗口。特征浏览窗口能够清楚显示轮廓特性的查找结果，如一个序列中某一部分序列所包含的 GC 含量。特征浏览窗口还能对一个序列中几个不同的部分进行比较，或对几个不同的序列中同样的部分进行比较。Composition Analysis 能以表格的形式显示序列中核苷酸或氨基酸残基的数目和百分比组成。利用上述的浏览窗口可进行大量的分析，通过 Project View

可以很容易进入这些浏览窗口。有些浏览窗口在后面的部分还将会详细介绍。

3.5 输入和输出序列

3.5.1 将序列输入到项目中

Omiga 的基本任务之一是序列输入，显然导入一个序列比用手输入序列更容易且不易发生错误。Omiga 支持多种序列格式，如 ASCII、EMBL、FASTA、GCG、GenBank、PC-Gene 和 SwissProt，输入的序列被转换成 Omiga 格式。Omiga 格式包含有一些附加的序列所有的特征和信息，如编码区、转录起始点、终止子和多聚腺苷酸信号等。在这个新输入的序列上，许多特征可以仅根据原始序列而被识别，但是一旦这些特征已经注释则常常是既有用又及时的。此外，来自 GenBank 序列的许多特征不是基于原始序列而是基于实验序列。因此，转录起始点，或内含子-外显子边界序列能够通过基因组 DNA 和 cDNA 的比较而检测到。这些元件并不是总能够通过这种简单的方法在仅有的序列中进行检测，故应用 Omiga 也可能不能检测到这些元件。例如，从 GenBank 导入一个大的核苷酸序列，如人的精蛋白基因簇(cluster)，则也会导入该序列的已注释特性，包括 TATA 框和 CAAT 框、转录起始点、外显子、起始密码子和终止密码子、多聚腺苷酸信号和外显子-内含子边界序列。虽然 Omiga 能够识别 ATG 或 TATA 基序，当一个特定基因确实有启动子和转录起始点时，只有实验性的证据是确定的。Omiga 也能够导入重复元件数据，这些元件由导入的原始序列所提供。显然，这是 Omiga 相当重要的一个功能。

3.5.2 输出序列

同输入序列一样，Omiga 能够以 ASCII、EMBL、FASTA 和 GenBank 格式输出序列，以及序列的特征信息。当将序列提交到不同的数据库时，这项功能被证明是十分有用的。因此，在实验室中获得的序列数据可采用 Omiga 提取信息。因而序列数据和采用 Omiga 识别到的所有的该序列的特性能够以 GenBank 或 EMBL 格式输出，提交到相应的数据库。GenBank 提供在线提交序列的方式，通过 Bank-it 界面提交序列。通过 Omiga 输出序列能加速序列提交过程。最后，序列及其所有附加的特征能够在两个或多个项目中转移。因此，在一个经常使用的载体序列中识别目的特性的操作只需进行一次，然后信息就可转移到另外的项目中或另外的用户。

3.6 序列比对

3.6.1 执行比对

Omiga 版本 1 采用 Clustal W 算法进行多序列比对。要使用 Omiga 进行比对，

先在 Project View(项目浏览)窗口中的树式面板中选取两个或多个序列。在工具栏中选中 Calculate(计算)选项，从显示的下拉菜单中选中 Align Sequence(比对序列)选项。Omiga 允许的核苷酸或氨基酸序列长度长达 10 000，序列数为 500 个序列。正如我们前面所讨论的，当进行过大数量的序列比对时，可使 CPU 负荷过重，甚至导致死机，因此建议使用较少的序列数。

Omiga 能够以两种方式产生比对。从 Project View 窗口中选取两个或多个序列进行比对，或将序列加入到 Alignment View(比对浏览)窗口中已经存在的比对序列中。这与 Clustal W 的界面很相似，该窗口也有两种操作方法。虽然添加序列到已经存在的比对中是典型的比较快的方式，但重新比对所有的序列(包括新添加的序列)却是常用的更为准确的比对方式。这是因为添加一个序列到已经存在的比对中，原来所有的比对序列仅作为一个序列(具有同一性的序列)，因此可能会丢失信息。然而，对于非常相似的序列，如来源亲缘关系非常近的物种的直向同源基因，添加序列到原来的比对中已经足够了，能够产生与较慢方法一致的结果。

Omiga 具有为进行比对而预设的许多参数，这些参数可根据用户的要求而改变，几组特殊的参数可保存为比对方法。这样，可以按照所希望的来选择空位罚分(gap penalty)、加权错误配对 (weighted mis-match) 和发散序列延迟(divergent sequence delay)，并作为当很多组序列进行比对时的使用工具，而不必每次都对比对参数进行定制，Omiga 也可以对评分矩阵(scoring matrix)进行选择，包括 BLOSUM、MD、PAM，并可以提供蛋白质序列比对的一致性(identity)，在文献(参考文献[2]和其他的参考文献)中也可以发现 Clustal W 和上面提到的各种评分矩阵特异性的额外信息，在此章中不做讨论(见第 11 章)。

3.6.2 编辑比对

除了进行比对外，Alignment View 窗口还可进行比对的编辑，添加注释。也可对比对的序列进行分组，如可自动地将一个序列的改动反映到组内的另一个序列中。当对比对结果不太满意时，需要在参与比对的多个序列的同一位置引入一段间隙时，这个功能就显得十分有用了。也可以将一个序列固定在比对窗口的上方，而不管窗口如何滚动。当查看一个序列与其他的序列一个一个进行比对时，这个功能就显得十分有用了。在比对的每一行的下方，Omiga 能够显示单个目的序列所具有的相似性或一致性的序列。在所有序列中具有一致性或相似性的序列用冒号标记，而与一致序列(consensus)或相似性很少的或有非常保守的差异位置则用句点标记。因此，在成对(pairwise)多肽序列比对中，赖氨酸和精氨酸之间用句点表示，因为它们都是碱性残基，尽管它们的核苷酸密码子有很大的不同，如赖氨酸为 AAG，精氨酸为 CGC。同样，酪氨酸和天冬氨酸残基(密码子分别为 TAY 和 GAY)尽管它们的化学性质十分不同，但由于具有保守的核苷酸编

码，因此也用句点表示。最后，在比对浏览窗口中，可以对比对的外观进行修改，这包括使用用户颜色方案、选择项目框或保守区域序列的阴影方式。这项功能以及彩色打印的功能，能够为用户提供核苷酸及多肽序列比对结果可视化信息文件(图 3.3)。

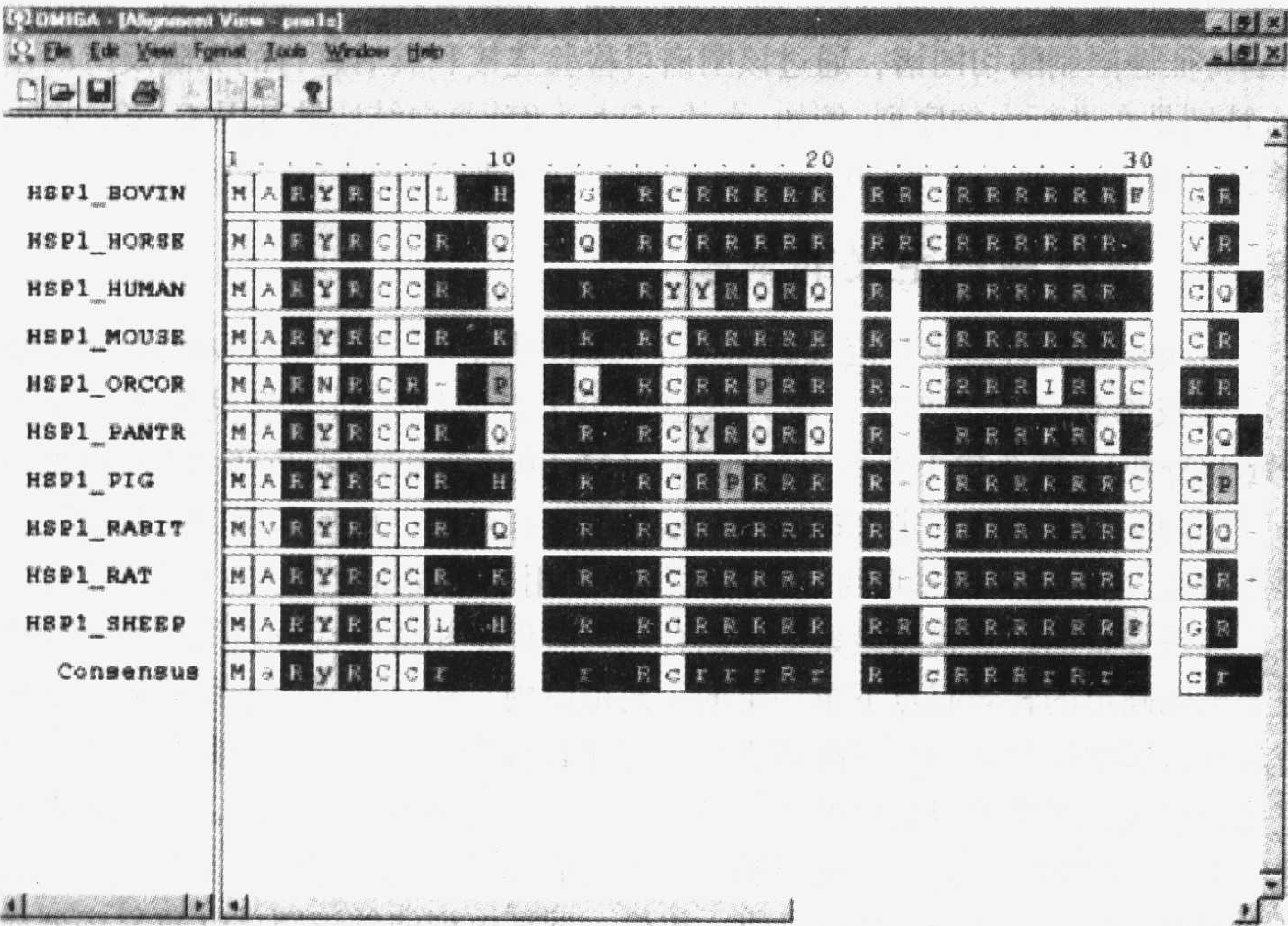


图 3.3　在 Alignment View(比对浏览)窗口中，比对结果能用各种颜色显示，这里用灰色阴影来显示

3.7　序列基序查找

3.7.1　核酸酶和蛋白酶位点的查找

尽管数据库提供的序列的特征及注释能够随着序列本身输入到 Omiga 中，但常常需要查找或添加序列的其他特征和注释，这或许是像 Omiga 这样的分析工具所具有的主要功能。用户可通过 Omiga 分别查找核苷酸序列中的限制性位点和多肽序列中蛋白水解位点。当查找限制性内切核酸酶酶切位点时，或是查找蛋白水解位点时，用户可以在序列中依次查找每一个核酸内切酶或蛋白酶位点。另外，Omiga 包含有所有常见的限制性内切核酸酶位点的 REBASE 数据库和蛋白水解的 PABASE 数据库，这些数据库可用来查找序列所有的位点。在工具栏中，在查找特性产生的下拉菜单中选取 Restriction Site(限制酶切位点)或 Proteolytic Site(蛋白水

解位点)选项，可进行序列的核酸内切酶或蛋白酶位点查找。在查找浏览窗口中，产生和显示的输出结果能够进行筛滤，得到进行亚克隆的理想位点。在这些指南中，初学者可通过输入人β球蛋白结构域序列，及查找可用于将β球蛋白基因亚克隆入 pUC19 载体中的限制性内切核酸酶位点的过程演示来学习。这个新的 pUC-β 球蛋白结构能够以一个新的序列形式被存储，并能查找限制酶切位点，给出该序列的特征性限制酶切图谱，通过该图谱可检验连接和其后进行的克隆反应的正确性。特别是在进行大的序列，例如，大约 75kb 人β球蛋白结构域基因(GenBank locus HUMHBB)操作时，采用 Omiga 查找这样的位点要比肉眼简便容易得多。

3.7.2 查找用户定义的基序

Omiga 也能进行用户定义的基序序列的查找，用户可利用 Omiga 存储基序并将查找参数存于方案中，该方案也可用于其他的序列中查找基序。除了从 Search(查找)下拉菜单中选中 Nucleic Acid Motif(核酸基序)或 Protein Motif(蛋白质基序)选项外，查找用户定义的基序的操作与查找限制性位点和蛋白酶水解位点的操作相似。用户定义的基序可以是非常简单的用于查找所有的 ATG 三核苷酸的读框，或是复杂的，查找与激素受体结合位点有设定相似性百分点的片段，该片段位于预先确定的转录起点下游一定的距离范围内。在启动基序查找以后，会弹出信息框，提醒用户利用核苷酸或蛋白质基序查找参数盒设置查找参数。用户自己定义的查找方案能够以数据的形式被保存，以供将来使用，因此，不必每次都要重新定义查找参数。这样，就能够查找到与启动子、调控序列，及其他顺式元件具有很高一致性的片段。理论上讲，提供一同源序列能够识别 Alu 重复序列或其他重复元件，这需要在目的序列中能够对用户定义截取值的重复元件进行识别。所有的这些元件都被标记并且作为该序列的特征。查找基序另外的一个用处是查找 PCR 产物中限制酶切位点，通过查找或识别以前设计的引物序列，用户可限定只在两个引物之间查找限制酶切位点。这对于通过限制性图谱表明成功获得特异性扩增产物的人来说十分有用。当用户获得熟悉了 Omiga 的界面和能力后，对这个功能的其他应用就融会贯通了。

3.8 查找编码区

Omiga 1 的另一个功能是在核苷酸序列中查找推定的编码区。可通过从 Search(查找)下拉菜单中选取 Open Reading Frame(可读框)选项来进行查找可读框。当对大片段的基因组序列进行分析时，这项功能就显得特别有用了。正如在前面谈到的序列比对一样，当查找序列中可能的编码区时，可根据用户要求生成特定的方案。Omiga 1 带用 GENMOTIFS 数据库的内部拷贝。另外，用户也能够添加基因(与基因相关的基序)的序列特征。例如，在大的基因组序列内，不均匀位置的碱基偏爱

性可准确地作为表达片段的指示[3]。但是将它归纳成与基因相关的基序会很困难，可以想像，其他的基序也可能通过这个编码序列的特征被查出。显然，查找编码区域在 Omiga 中很复杂，可采用不同权重的多个基序，也就是使 Omiga 在序列中查找具有统计学意义的基因相关基序簇(cluster)。但这个方法可能比采用神经网络方法的查找工具，如 GRAIL[4]的效果差一些。

3.9 引物查找

Omiga 增加的另一个很好的功能是引物设计。可以根据用户定义的标准设计测序或 PCR 引物。首先在树式面板(tree panel)浏览窗口中选取一个序列，然后从 Search(查找)下拉菜单中选取 PCR Primer Pairs(PCR 引物)或 Sequencing Primers(测序引物)选项，就可进行引物设计了。Omiga 提供了默认的参数，用户也可自己定义查找参数，并将其存为另一个方案。当查找 PCR 引物时，用户定义的参数包括引物的长度、GC 含量、解链温度、盐浓度以及引物浓度。用户可以在模板序列中指定特定的区域查找引物，或不在该特定区域查找引物。用户可以指定一个 3′夹和含许多多义核苷酸的简并引物。引物设计的另外一个特点是能够去除多个扩增末端，即如果在一个很大的模板序列中，一个 5′端引物与多个潜在的 3′端引物相适合，则只产生最短的扩增产物，因此减少由其他相符的引物对该引物的干扰。然而，让用户设计与一条已经定义好的引物相匹配的引物，在 Omiga 好像没有此功能。例如，想设计一对引物，该引物对能够扩增跨越内含子的片段(可区别基因组 DNA 和 cDNA)，然后另外设计一个附加的 3′端引物，使该引物与最初引物对中 5′端引物相匹配。这个附加的 3′端引物应当设计在内含子中，可以鉴定前 mRNA，或在不同的外显子中显示不同的剪接。在其他的引物设计工具中，为 PCR 扩增查找一个与预选好的引物相匹配的引物，这是一个很常见和有用的功能。

当查找 PCR 引物对时，Omiga 能够识别三种类型的引物与引物之间的相互作用，包括引物-引物退火、引物自身退火和引物与模板退火。用户可以指定 Omiga 处理这三种类型的退火。引物与引物之间的退火包括引物二聚体的形成，如果形成非常稳定的引物二聚体，特别是在引物的 3′端，将会产生很少或没有模板的扩增产物的结果。在每一单个引物内部引物自身退火能够形成发夹结构，与引物二聚体一样，如果某引物内部很大的片段形成发夹结构，也将有很少或没有扩增产物产生。大量引物由于用于形成发夹结构而被封闭，因而不能与模板退火。最后，Omiga 也能够查找引物与模板序列的非特异性退火。这对于避免产生包含过于简单元件的引物很重要，那些包含简单元件的引物可能与模板中的重复区域退火。Omiga 缺乏的另一个引物设计功能是针对附加的数据库序列查找引物。虽然将引物与大数据库相比较很费时，但这也是一个很有用的功能。例如，当在一个简单生物的基因组中设计一对引物以扩增某区域时，而且这个基因组的序列是已知的，

可以通过比较引物与这个生物的整个基因组，避免出现假引发。在基因组时代，这种查找是可能的。同样，如果在公共数据库的短序列中设计引物，可能需要对已知的重复元件的数据库扫描这些引物(即查看引物是否为这些重复元件)。而这样的序列在编码序列中很少，因此，不可能通过运行 Omiga 的引物-模板退火比较查找到，即使出现与 Alu 或 Line 轻微相似的元件，也常常会导致出现不适当的引发，以及目的产物的扩增很差。

在 Omiga 中除了不必考虑引物-引物相互作用这一项以外，测序引物的查找与 PCR 引物对的查找也很相似。当然测序引物的某些特点与 PCR 引物还是有所不同的，因此，Omiga 采用了不同的默认设置。正如 PCR 引物查找一样，用户可以对这些值进行编辑，并形成方案。因此，可以利用这个特点设计测序用途外的单个引物。例如，用 poly-T 引物不足以扩增 cDNA 近 5′端极长的信息区域时，用户可设计一个引物以扩增该区域。当为 5′和 3′RACE 实验设计引物时，这项功能就十分有用了。

查找完引物及引物对后，结果可以以图形或表格的格式显示，以图的格式显示用 map 按钮，以表格的格式显示用 table 按钮。在图的格式中，以一个图显示所有合适的引物，或所有相匹配的引物。另外，如果没有查找到合适的引物对，Omiga 能够让用户分析失败的原因。通过这个途径，可以改变个别的查找参数，直到查找到合适的引物。最后，正如 Omiga 的许多其他特征一样，制定的单用户的引物查找方案可被应用于多个项目，以及被多个用户使用。因此，实验室的所有成员都能共享这个最佳引物查找方案。

3.10 注释和注解

Omiga 1.1 版本可用于大多数分子生物学实验室的辅助实验设计。Omiga 的操作界面学习和使用都非常简单。Omiga 最大的缺点可能是做某些分析时有些繁杂。对某些分析，如 Omiga 1 提供的第三个指南的一个细节，需要同时打开很多窗口，这使得操作变得笨拙。但由于用户要使用大量的功能，因此是不可避免的。采用 Omiga 设计一个完整的实验是完全可能的，从在 cDNA 文库中扩增目的序列到限制酶切扩增产物的克隆再到鉴定目的构建产物成功分离。可以理解，由于要处理大量的信息，因此，有时导致计算机屏幕上出现过多窗口。尽管如此，Omiga 1 仍是一个很有用的工具，是大多数分子生物学实验室必需的基于计算机的序列分析工具。

在 Omiga 将来的版本中应当包含这样的一个有用的特性，那就是能够输入由 Clustal W 或其他的比对工具产生的比对结果。尽管 Clustal W 是一个非常有用和流行的比对工具，但还有一些其他的比对工具，其中有些工具也比较受欢迎。还有一些其他的特点，如序列编辑/片段重叠接头查找功能将有助于提高 Omiga 的效

用。显然，基因组测序中心在大型主机上应用软件编辑序列，但其他的 PC 和 Mac 软件工具也适合编辑序列。牛津大学分子组(Omiga 的制作者)最近开发了 GCG。GCG 是当前应用最广泛，分析序列最全面的一种工具，将来 Omiga 也能够支持 GCG。如果这两个工具能够完全兼容，那将是非常理想的。这样研究者能够应用基于强大的计算机主机的 GCG 程序编辑测序实验的信息，同时也具备像 Omiga 这样基于 PC 机的程序进行简单和常规的分析的能力。引物查找方案是 Omiga 1 具有的另一项很好的附加功能，但正如前面讨论过的，还缺乏一些与其他普通的引物设计软件相似的附加功能。

(李　鹏　译)

参 考 文 献

[1] Oxford Molecular Group. (1997) *Omiga 1.0 User Guide.* Oxford Molecular Ltd., Oxford, England.

[2] Thompson, J. D., Higgins, D. G., and Gibson, T. J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucl. Acids Res.* **22,** 4673-4680.

[3] Kramer, J. A., Adams, M. D., Singh, G. B., Doggett, N. A., and Krawetz, S. A. (1998) Extended analysis of the region encompassing the PRM1→ PRM2 → TNP2 domain: genomic organization, evolution and gene identification. *J. Exp. Zool.* **282,** 1-2, 245-253.

[4] Uberbacher, E. C. and Mural, R. J. (1991) Locating protein-coding regions in human DNA sequences by a multiple sensor-neural network approach. *Proc. Natl. Acad. Sci. USA* **88,**11 261-11 265.

4 MacVector：Macintosh 计算机集成序列分析软件

Promila A. Rastogi

4.1 引言

对任何一个研究人员来说，无论他正从事基因组计划研究，还是对目的基因进行克隆或定性研究，操纵、分析和注释序列数据的能力变得日益重要。到目前为止，对几千或几百万个碱基手动有效地进行以上操作是十分困难的，甚至几乎是不可能的，因此能够进行序列分析的计算机程序在从事分子生物学的实验室得到越来越普遍的应用。MacVector 是由牛津分子研究组(Campbell CA)开发的一种计算机软件包，是一种在 Mac 机上运行集成的复杂的序列分析程序。它提供所有最普遍使用的核酸和蛋白质分析工具，而且提供通过因特网进入美国国家生物技术信息中心(NCBI)公开的 Entrez 数据库。到写作本文时，MacVector 的版本为 6.5。MacVector 还有一块用于重叠片段接头组装的模块，称为 AssemblyLIGN，在这里我们将不讨论 AssemblyLIGN。

4.2 材料

MacVector 6.5 需在 MacOS 7.1 或更高的版本上运行，需要一个光驱和 18M 的磁盘空间。所需的最小 RAM 内存为 4M，但建议至少使用 6M 的 RAM。要进入 NCBI 数据库，需要连接到因特网上。为了最好地利用信息，建议使用彩色显示器和 Mac 兼容的打印机。用户指南[1]以书的形式提供，并以一个 PDF 文件存放在 MacVector 6.5 的光盘上。

作为一个商品程序，MacVector 具有防止拷贝的功能。可以通过硬件或软件来防止拷贝。硬件防拷贝装置称为 EvE 密钥，它与 Mac 的苹果桌面总线(ADB)相连，该装置可以与计算机的键盘或鼠标进行菊瓣式连接。在将 EvE 往 ADB 上装卸时，必须先关掉计算机。软件防拷贝的方法是 Sassafras 软件™ (Hanover, NH) 的 Keyserver™。Keyserver 用于防止在同一软件许可证内进行多份拷贝，它安装在网络中的一台计算机上。MacVector 安装在每个用户的机器上，它通过 AppleTalk 与 Keyserver 相连，以防止非法拷贝。Keyserver 在编程时已输入许可的用户数(x)，

若第(x＋1)个用户试图登录到他/她的 Keyserver 上，必须等到其他用户中的一个退出 MacVector 程序，因为此时 Keyserver 的防止拷贝功能才停止工作。

4.3 方法

4.3.1 普通程序界面

MacVector 遵循 Macintosh 的界面准则，所提供的界面易于使用并为任何 Macintosh 用户所熟悉。程序的菜单条有 6 项：File(文件)、Edit(编辑)、Options(选项)、Analyze(分析)、Database(数据库)和 Windows(窗口)。File、Edit 和 Windows 菜单与其他 Macintosh 程序的菜单相似。Options 菜单提供了三条命令，可用于改变在分析结果窗口中信息呈现的方式(注释显示格式、比对显示格式和默认符号)。用户通过选择 Modify Genetic Code(修改遗传密码)命令，可以修改或创建在所有翻译中使用的遗传密码分配(assignment)，或通过使用与 NCBI Entrez 数据库相连的 Make Codon Bias Table 命令创建一种微生物的密码子偏爱表。Analyze 菜单列举了 MacVector 对核酸和蛋白质序列能够进行的各种分析功能。Database 菜单包括有关数据库搜索的命令，例如，BLAST 搜索、Entrez 浏览和要查询的序列与存在当地的序列数据库之间比对。

序列是 MacVector 的中心对象，在对一个序列进行分析之前，必须将序列窗口激活。程序是高度交互式的。Analyze 菜单中的命令是否可以使用取决于所激活的序列的类型。当一个核酸序列被激活时，所有蛋白质的分析功能呈灰色，反之亦然。大部分分析的结果都保存起来，这样用户可以应用各种过滤器来反复考察不同结果的子集，而不必重做所有的分析。许多分析也提供大量不同的考察结果的方法：以表格形式、图形形式和注释序列的形式。图形结果的局部区域能放大，用户得以观察到更多的细节，结果可以保存为 Text 和 PICT(图形结果)文件，这样就能在 Word 程序或绘图程序中复审或编辑。

当用户启动一个分析时，就会出现与此分析相对应的对话框，在对话框中，用户既能键入他们选择的参数，也能选择默认的参数执行分析。用户提交的分析立刻就能完成，MacVector 将结果以对话框的形式呈现。且大多数分析结果能以不止一种形式呈现出来，结果呈现出来之前，用户可以在对话框中选择各种结果的形式，也可以通过改变任一可能的选项来过滤结果。一旦选定结果的形式，MacVector 就将其显现出来。

4.3.2 输入和编辑数据

MacVector 支持大多数常见的序列格式，原始的 MacVector 序列文件类型为二进制文件，其他的文件格式有文本格式，并包括 GenBank、GCG、IG-Suite、

CODATA(包括 NBRF PIR、EMBL、SWISS-PORT、FASTA)和 line(只指序列)。这些格式的序列文件在 MacVector 能用 File | Open 打开，然后用 File | Save 命令以 MacVector 的格式保存；也能手动输入建立序列，从紧靠 File | New 命令的下拉式菜单中选择序列类型，从而打开一个新的序列窗口，像 Word 程序那样用键盘或数字键输入实际残基即可。当序列输入后，校对员进行检查。蛋白质序列也可通过 Analyze | Translation 命令翻译部分或整个核酸序列来创建。核酸序列也能类似地建立，通过 Analyze | Reverse Translation 命令来逆翻译部分或整个蛋白序列。

序列窗口有按钮将其链接至序列的 Features Table、Annotations Table、Annotated Sequence View 和 Feature Map(图 4.1)。其中 Features Table 和 Annotations Table 既可分析原始的 GenBank 序列文件中的信息，也可分析用户输入的序列中的信息。Annotated Sequence View 是陈述序列所注释的特性文件。Feature Map 是陈述包含在 Features Table 中信息的专业图形。

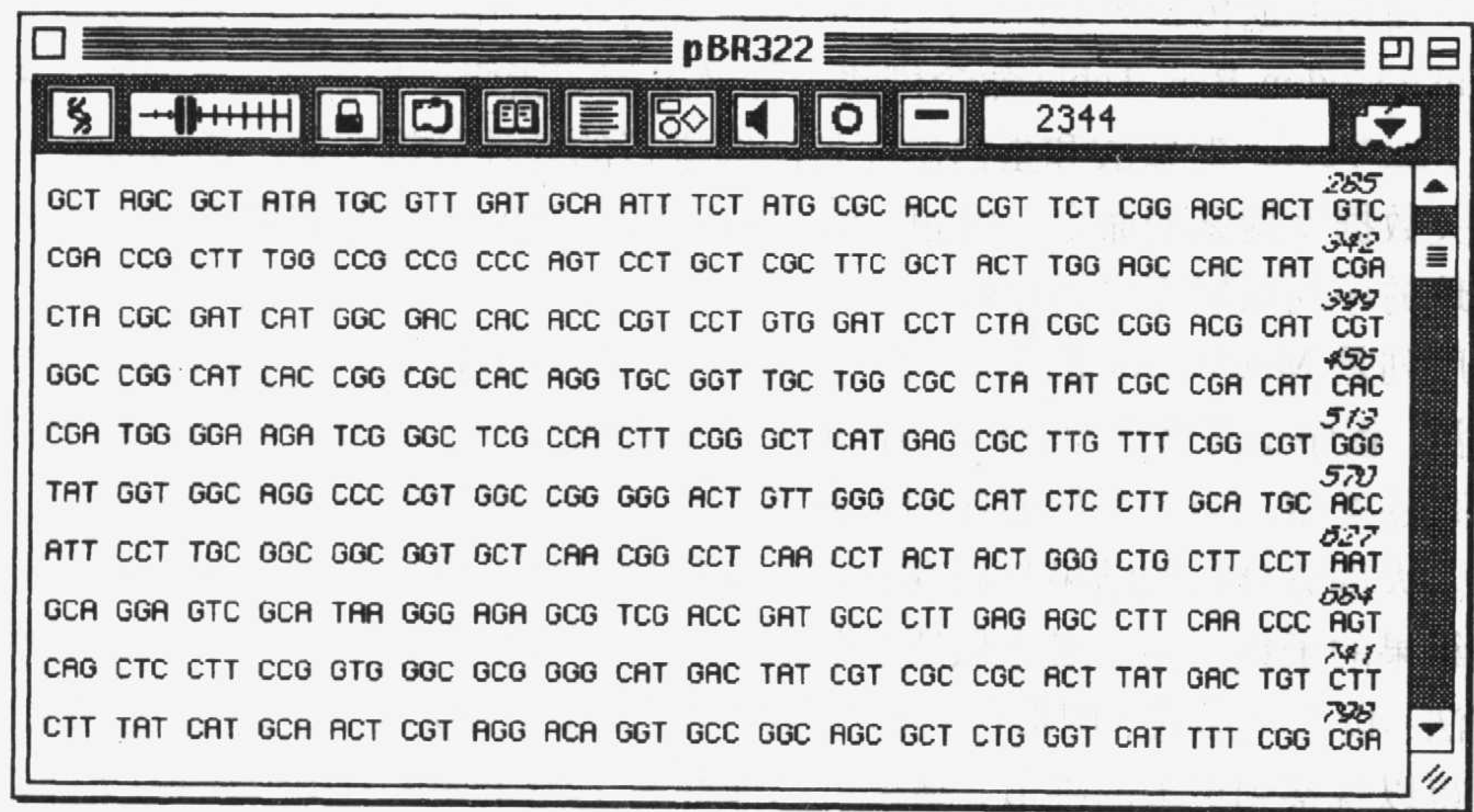


图 4.1　pBR322 序列窗口

MacVector 带有各种数据文件类型，用于它所提供的许多分析中，它们都是二进制格式，可以在 MacVector 内进行编辑。这些文件包括限制酶、蛋白水解剂、核酸和蛋白质的子序列(基序)，以及对核酸和蛋白质序列比较的评分矩阵(scoring matrix)。在 File | New 命令的下拉菜单选择文件类型就能为这些数据文件建立新窗口。程序提供大量生物体的碱基偏爱密码子表，它们能在 MacVector 中建立，但不能由用户修改。

4.3.3 输出序列

序列能在 MacVector 中以下列形式输出：GenBank、GCG、FASTA、IG Suite、Staden 和文本(ASCII)，利用 File | Save As 命令就可以输出序列。结果文件将以文

本文件保存在用户选定的地方。双击以这种方式建立的文件就能登录 MacVector，以 MacVector 格式打开序列。以 GenBank 或 GCG 格式输出序列，必须确保与序列有关的特性和注释随序列一起输出。

4.3.4　定制特征图谱和结果图谱

MacVector 6.5 版本比以前的软件版本在图形方面更佳，特征图谱和结果图谱都可以定制：用户能选择颜色、模式、单个特征和结果的风格，以及在特征图谱和结果图谱上标注和标题中所用的字体和文本风格。用户也能决定哪种类型的特征和结果显示于图谱上(图 4.2)。

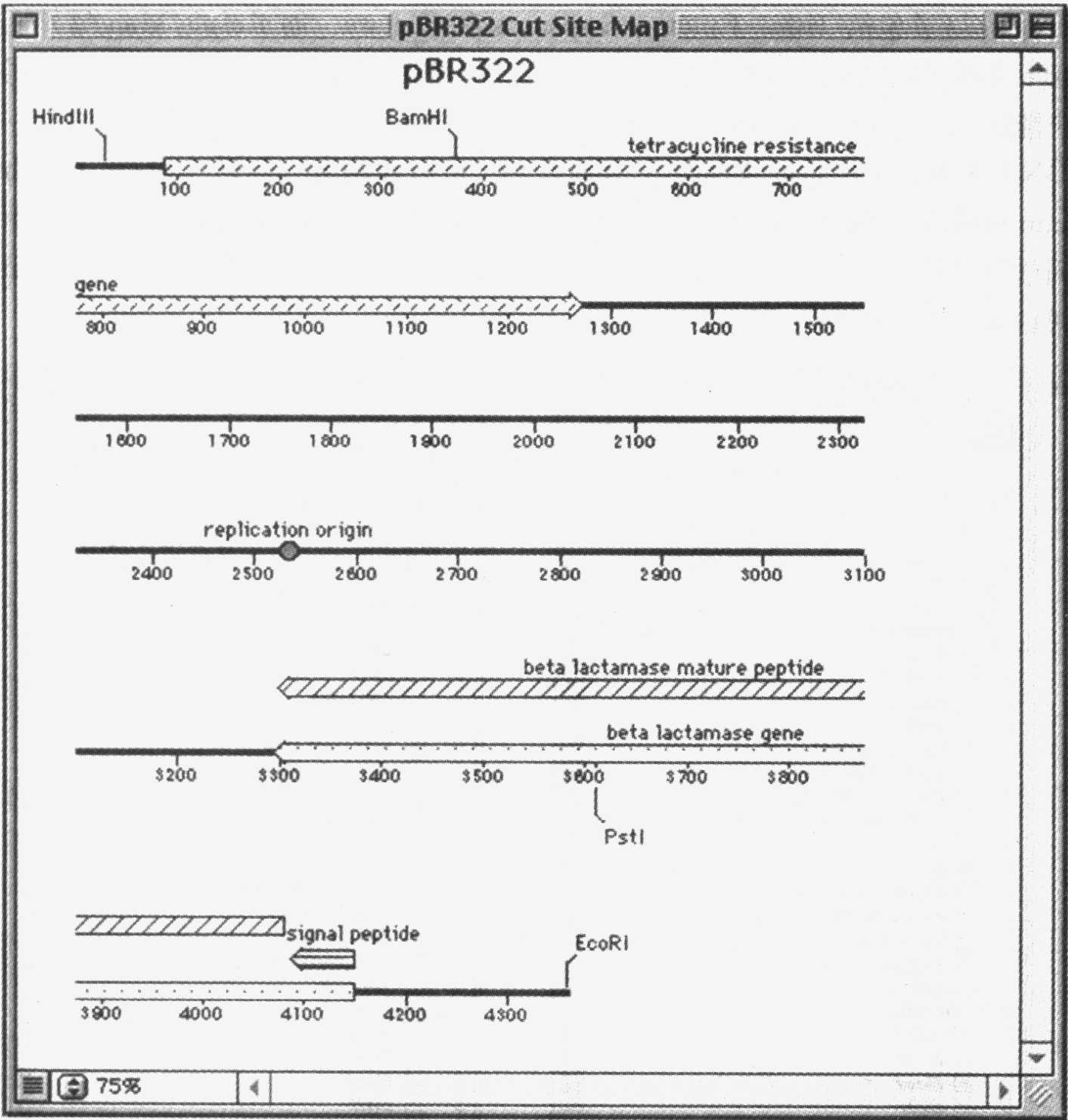


图 4.2　pBR322 限制性图谱

在此图谱上，用户已定制了特征的外观

Options | Default Symbols 命令允许用户为特征图谱和结果图谱设置共用的符号。只有当序列窗口或结果窗口未在桌面上激活时，此菜单的选项才有用；若已激活，则选择各菜单选项时，必须按住 Option 键。默认字符窗口在左上角有一下拉菜单，提供 Title、Feature、Result、Ruler 和 Sequence 选项。如图 4.2 所示，用户能选择字体、风格、颜色，以及标尺、序列的位置和标题，也能为每个单独的特征类型和结果类型设定默认符号，这包括风格(外形)、颜色、填充模式、特征和结果的位置、特征和结果标注的位置以及标注文本的字体、风格和颜色。当序列窗口激活时，单个序列的符号能以类似的方式设定，即选择 Options | Symbols for〈sequence_name〉。

当序列的 Feature Map 或 Result Map 为激活窗口时，Graphics Palette 对话框也会显示。对话框的滚动列表中列出了专用图谱元件，此元件能被编辑。在 Feature Map 中，目录包括 Title、Ruler 、Sequence 和 Feature，而 Result Map 另外还有 Result 选项。Graphics Palette(图 4.3)允许用户随意打开或关闭专用图谱元件，例如，用户能让某一个特征在图谱上不显示出来或只显示某一特定的限制酶切位点。对于 DNA 序列，用户通过选择线状/环状按钮，可以在两者之间转换显示。Graphics Palette 也能允许用户调控残基的密度(每一英寸、厘米上碱基的数目，或线型图的排列或环状图的半径)，线型显示的长度和图谱显示序列的范围。另外，它的 Filter 按钮也能为 Result Map 进行结果审查。

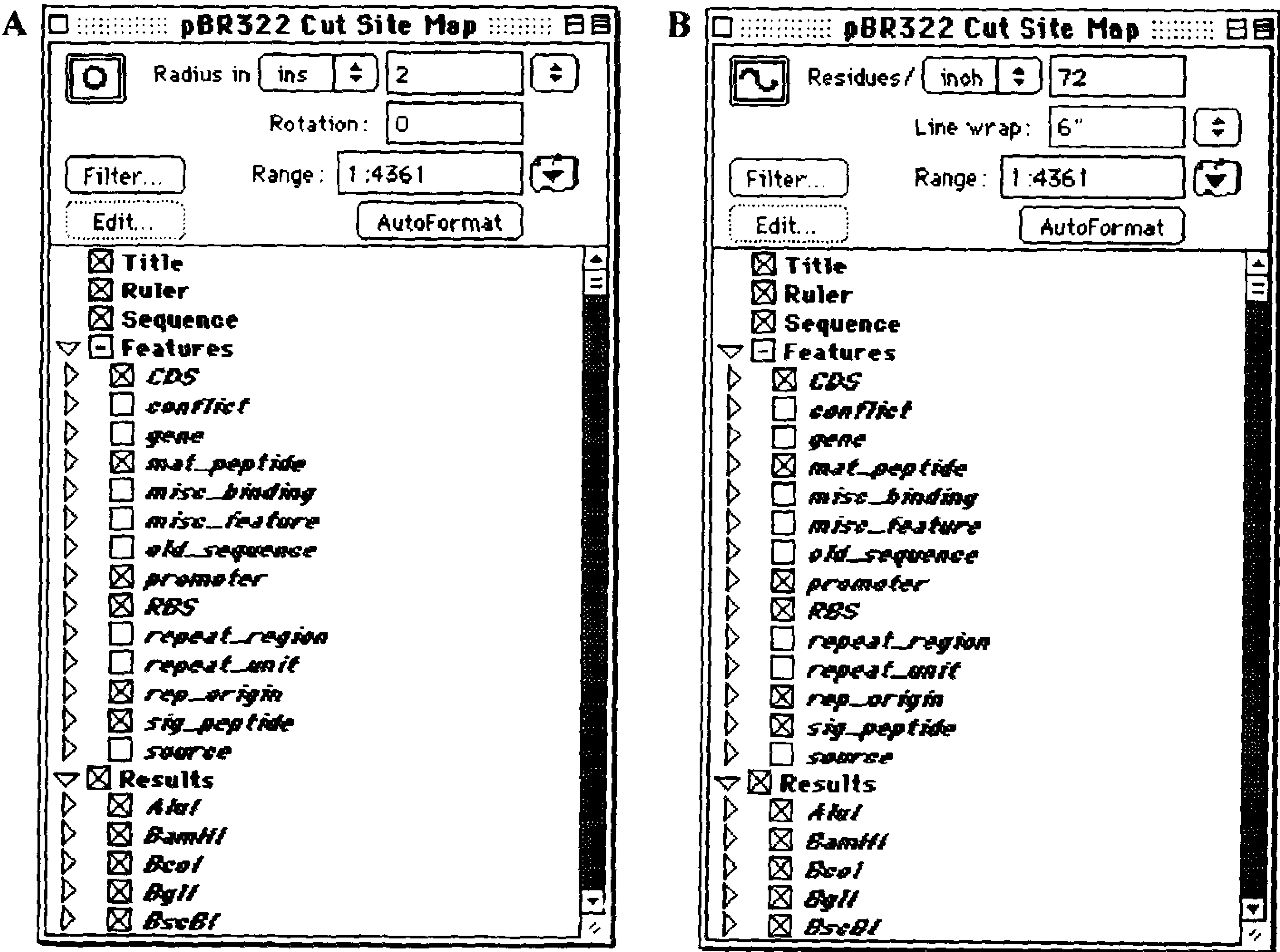


图 4.3 Graphic Palette 下 pBR322 限制性图谱的环状(A)和线状(B)形式

用户可以在图谱元件的 Check 框中打开或关掉此元件

4.3.5 DNA 分析

4.3.5.1 限制性分析和基序分析

MacVector 带有由几个限制酶文件组成的名为 REBASE 的限制酶数据库，这些限制酶文件列出了不同商家的酶。用 File | Open 命令打开这些文件：每一行列出的分别是名称、酶切位点和某一酶的所有已知同裂酶(图 4.4)。点击某个酶，在它名称的左边就会出现一记号，从而便选择了此酶做限制性分析。选择 Analyze | Restriction Enzyme 命令，弹出限制酶分析对话框，点击 Enzyme File 按钮，就选择了一个限制酶文件用来分析。在 Search using 的下拉菜单中用户可以选择限制酶文件中的所有酶或以前选定的某个酶。

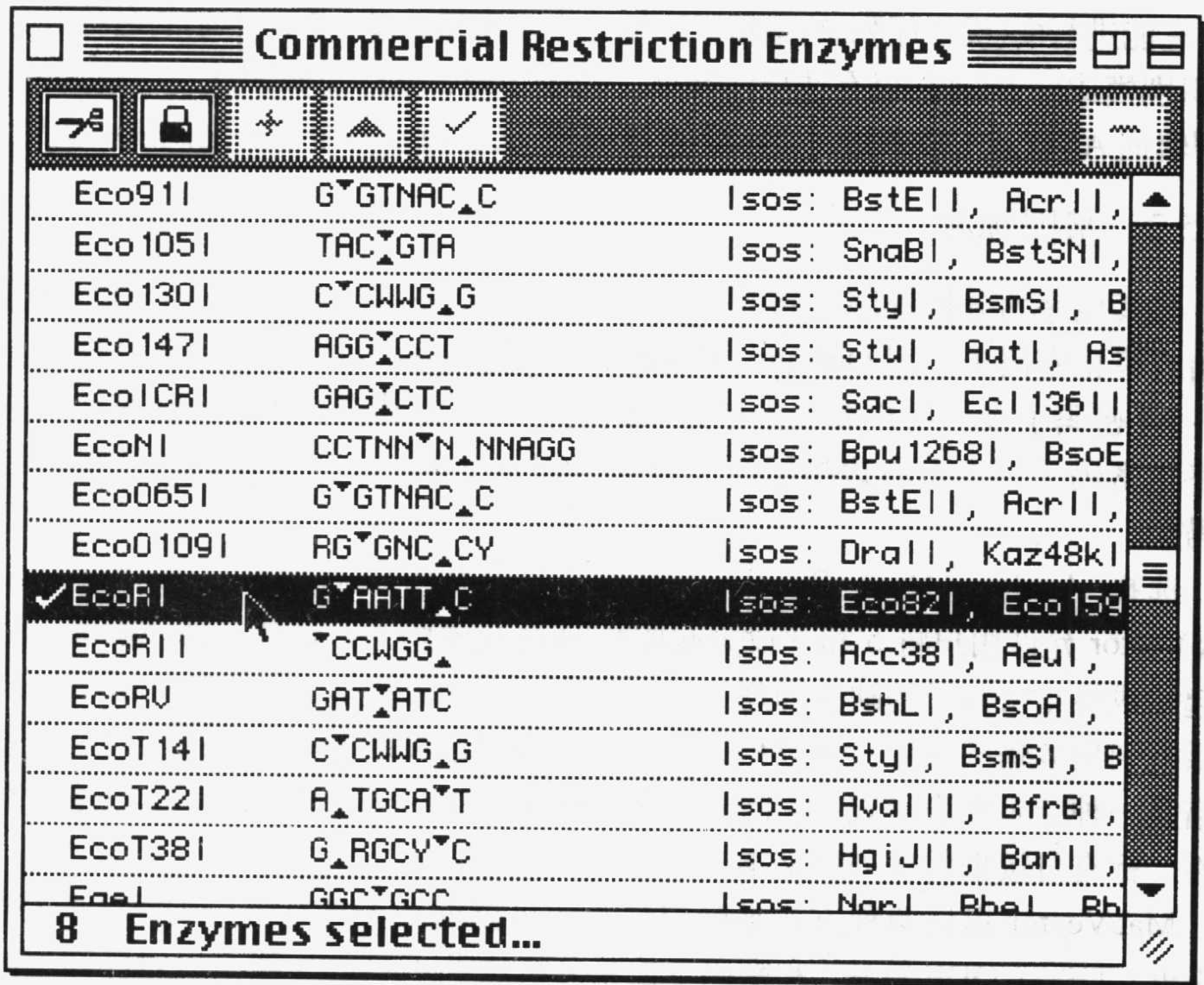


图 4.4　一个 MacVector 中的限制酶文件

*Eco*R I 是所选定的 8 个酶中的一个，由靠近该酶名称的对号标明

当分析完成时，结果对话框列出输出选项：存在位点的限制酶的列表、无位点的限制酶的列表、限制酶切图谱、注释序列的酶切位点和酶切片段的预测。限制性图谱可以定制成包含特征的形式，可以对结果进行过滤，这样，只显示酶切位点的一个子集。通过点击 Graphics Palette 中的 circular/linear 按钮，图谱就能在

环形和线形之间互换。用单酶或双酶消化的片段也能预测。

在 MacVector 中，基序指子序列，MacVector 带有 6 个核酸子序列文件，每个文件都可以用 File | Open 命令打开。在限制酶文件中，每个子序列文件的每行包括名称、识别位点和核酸子序列的附加信息。对子序列全部选定使其醒目，复制到写字板并粘贴成一新的核酸子序列文件(通过 File | New 命令打开并从下拉菜单中选择 Nucleic Subsequence 项)。保存此新的核酸子序列，然后用于核酸子序列分析。它类似于限制性分析，选择 Analyze | Nucleic Acid Subsequence 命令进行。

核酸子序列分析的结果对话框与限制性分析相似，结果选项列出存在的位点和不存在的位点列表、子序列图谱、注释序列的子序列和片段预测。最后一项是指给定序列所出现的区域之间计算出来的大小。

新的限制酶(或子序列)能添加到一个限制酶文件中，通过点击“＋”按钮实现，它能进行酶(或子序列)的编辑。这种编辑要求输入酶(或子序列)的名称和位点(或识别模式)。同时它也有选项注释框。若子序列，基序总数多达三个也可输入，用逻辑词 AND 或 OR 来指定,若 AND 已被使用,也可在各个部分之间隔一段距离。

4.3.5.2 基因预测

与其他序列分析程序不同的是，MacVector 允许用户自定义某个密码子作为起始或终止密码子。用 Options | Modify Genetic Codes 命令打开一对话框，它包含一 64 密码子表。用户能点击单元格内的密码子使其变亮，就可将其设为起始或终止密码子。点击右边指示箭头就设为起始密码子，点击左边的则设为终止密码子。

选择 Analyze | Open Reading Frames 命令就可以启动可读框分析。由于 MacVector 允许用户将 5′和 3′设为起始和/或终止密码子，这样，就能确保不会因为起始或终止密码子的缺失而忽视超出序列末端的可读框。当分析完成后，结果对话框列出各输出选项：可读框列表、可读框图谱和标有可读框的序列。可读框图谱是一种结果图表形式，且能显示在序列 Features Table 中出现的部分或所有的特性。标有可读框的序列以文本格式显示出可读框。

MacVector 提供两种方式来确定预测的可读框是否具有蛋白编码区的特性：Fickett's TESTCODE 算法[2]和密码子偏爱分析法[3]。Fickett's 法是以对已知编码区和非编码区进行统计分析为基础的，且能用来预测 200bp 以上的 DNA 序列的编码区。选择 Analyze | Open Reading Frames 命令，再选择 Fickett's method 复选框就能用 Fickett's method 法来预测可读框。

密码子偏爱分析法是基于这样一个原理：根据某些生物体所利用的 tRNAs 的种类，相应地推导出对密码子的使用有偏爱性，非编码区无密码子偏爱性；表达水平较低的蛋白其编码区密码子的分布类似于相应 tRNAs 的水平；而表达水平高的蛋白对相应的含量丰富的 tRNAs 显出高偏爱性。在应用密码子偏爱分析法前，

必须获得生物体的密码子偏爱表。若 MacVector 不提供此表，可以用 Options 选项中的 Make Codon Bias Table 命令，从 NCBI 中的 Entrez 数据库获得。打开 Analyze 选项中的 Codon Preference Plot 命令就能进行实际的密码子偏爱分析操作，并且对每个读框结果进行列表。密码子偏爱表中的突出点为某个密码子偏爱区域的指示(可能编码高表达的蛋白质)。

4.3.5.3 引物分析

MacVector 能根据用户选择的参数来设计某给定模板的 PCR 引物[4]。用户需为 MacVector 所设计的引物提供一个标准，程序对序列进行扫描之后，列出一系列的引物，对不符合用户标准的引物进行删除。MacVector 也能检测引物与用于 PCR 的模板序列之间的结合，这样，用户输入引物序列，程序对其进行分析，并显示出引物的信息，用户就能根据分析的结果决定它是否用于实验当中。同样，MacVector 也能预测用于模板测序的引物和检测用户所选的结合模板的引物能否作为测序引物。PCR 引物分析功能可以预测或检测引物对，而测序引物分析功能只能预测或检测单条引物。

选 Analyze 菜单中 Primers | PCR Primer Pairs 命令启动 PCR 引物对预测；测序引物的预测则选 Analyze 菜单中的 Primers | Sequencing Primers/Probes 命令即可。初始对话框可让用户指定目的区域；对于预测 PCR 引物对，是以序列大小和侧翼序列为基础的，用户能选定的其他参数包括引物-模板双链体的熔解温度，引物的 G+C 含量，引物要求的长度以及引物是否需要 G-C 锚。当引物自身形成发夹结构或引物之间形成双链时，用户可以指定引物连续结合的最大数目，以避免形成过多的二级结构，也可避免引物结合到其他的位点。对于 PCR 引物对，程序可以进行每条引物的 3′末端与反应扩增产物的比较。对于测序引物，比较的是每条引物的 3′末端和整个序列或序列某个指定的区域。

结果对话框列出的是一系列统计结果，包括检测的引物数目，接受的引物数目和不接受的引物存在的原因以及结果以表或图的形式列出，列出的 PCR 引物每对都提供 T_m 值、GC 含量信息。

选 Analyze | Primes | Test PCR Primer Pairs(或 Analyze | Primers | Test Sequencing Primers/Probes)命令就可以启动对用户所选定的引物进行检测。用户将一条或两条引物输入对话框(对于测序引物，只能输入一条)，用户点击 Apply 按钮，就列出引物的统计结果(图 4.5)，包括：T_m 值、引物的 GC 百分含量、退火引物是否适合模板、引物是否具有过多的二级结构信息。若是用户选定的引物对，另带有 PCR 的相关信息，包括产物大小、引物是否结合在产物的 3′端、T_m 值的差异，任何不利于 PCR 反应成功的信息都会用红色标示增强亮度。

ECUVRC PCR Characteristics

Primer 1: CATGGTCTGTGTAATGCGGAGAC　Parameters

length: 23, %GC: 52.2, Gs: 8, Cs: 4, ambiguous G or C: 0
1 ug of primer is equivalent to 400.4 pmole of ends
Primer does not form Self 3'-dimer
Primer does not form Hairpin
Primer does not form Self Duplex
Tm: 55.9 °C, (Of Primer itself)
Primer binds at position 2551 on the Top Strand (score 23)

Primer 2: GGTAACTATCTGTTGTCAGTA

length: 21, %GC: 38.1, Gs: 5, Cs: 3, ambiguous G or C: 0
1 ug of primer is equivalent to 447.7 pmole of ends
Primer does not form Self 3'-dimer
Primer does not form Hairpin
Primer does not form Self Duplex
Tm: 38.3 °C, (Of Primer itself)
Primer binds at position 4467 on the Bottom Strand (score 21)

pairing:

Major product size: 1917 bp
Primer pair does not form Duplex
Primer pair does not form 3'-dimer
Primer pair Tm difference is 17.6 °C

Region... 1 to 4549　Apply　Cancel　OK

图 4.5　用户定义引物的测试对话框

核酸杂交探针也能用于引物相同的模式来预测，若碱基序列未知，但氨基酸序列已知，使用反翻译功能，以获得杂交探针。MacVector 扫描整个翻译过来的 DNA 序列，寻找非简并区域，并输出一寡核苷酸目录，同时列出退火温度，G+C 含量和覆盖所有序列所需的简并寡核苷酸的个数。用户可以通过减少每个氨基酸的密码子数(基于生物体密码子偏爱性)来修改用在反翻译中的遗传密码，这就可以减少核酸探针的简并性。

4.3.6 蛋白质分析

4.3.6.1 蛋白酶和基序分析

蛋白质模式分析类似于核酸模式分析，蛋白水解酶和子序列分析分别采用蛋白水解酶文件和子序列(基序)文件。打开子序列文件，并点击文件名，就可选定它，且在序列左边出现一标记记号。蛋白水解酶的选定也类似于此，只是它打开的是由 MacVector 提供的蛋白水解酶文件。另外，与 DNA 核酸限制酶和蛋白子序列相似的是，点击“＋”就可在蛋白水解酶文件(或子序列文件)中添加一新的蛋白水解物(子序列)，点击“－”就可去除，“△”则是进行

修改。

蛋白水解酶文件含有化学和酶学水解试剂，也具有酶切位点和参考信息，通过 Analyze | Proteolytic Enzyme 命令来启动分析。分析对话框允许用户选择蛋白水解酶文件，以及程序是使用文件中的所有蛋白水解试剂，还是只使用用户选定的试剂。用户也可选择所要分析的序列区域。结果对话框允许用户选择输出选项：列出有酶解作用或者不具有酶解作用的蛋白水解试剂、一个水解图、带有酶切位点的文本格式的序列，以及预测的单酶或双酶消化后的片段。

蛋白子序列分析既可对选定的序列也可对所选子序列文件中所有的子序列进行分析。子序列自身由 2~3 部分组成，Analyze | Protein Subsequence 命令启动此分析，结果列出在序列中发现的子序列和未发现的子序列、子序列图，以及带有子序列标示的文本格式的序列。子序列图可以修改，以便显示出蛋白序列带有子序列的特征，结果可以进行过滤，可以只在图上显示所有子序列的集合。

4.3.6.2 蛋白质分析工具框

Protein Analysis Toolbox 是用来分析一给定蛋白质序列的大量特征，利用这些特征再推导蛋白质的三维结构信息和可能的功能信息。下述各种特性：二级结构、亲水性、疏水性、抗原性、柔性、表面均一性、两性和预测的跨膜螺旋，Toolbox 提供的算法不止一种，其他还有氨基酸组成、分子质量和等电点分析。所有这些方法是建立在已知结构和功能的蛋白质的数据基础上的。若分析一结构和功能未知的蛋白质，用户首先必须检测此新蛋白质是否与任一已知蛋白质相似，在对一新蛋白质的结构和功能做结论前，都应当综合 Protein Analysis Box 中算法提供的信息和生物化学和生物物理方面的数据。

二级结构预测有如下两种方法：Chou-Fasman 法[5]和 Robson-Garnier 法[6]。MacVector 软件并不能解决这两种方法之间的任何矛盾，用户也应该认识到每种方法只有 60%的正确率。因此，即使两种方法在序列某区域的构型预测上是一致的，也不能认为这种预测就完全正确。

Protein Analysis Toolbox 是通过 Analyze | Protein Analysis Toolbox 命令进入的，此对话框有 4 个小窗格(图 4.6)，第一格列出 Toolbox 中所有的算法：用户可以选择是列出还是用坐标标出每个结果，也可以选择列出 pI、分子质量和氨基酸组成。第二格让用户选择所要分析的序列区域。第三格让用户选择每种算法窗口的大小。当用户在第一格标示出某一算法时，第四格就出现此算法的简单描述。列出结果是以列表方式显示分析结果，每一纵列列出的是某算法的结果，每一行则代表序列中的氨基酸。坐标图是在一单独的窗口显示，用光标标亮某一区域就可使图放大，双击鼠标就可使图恢复到原来的尺寸。氨基酸组成、分子质量和 pI 在第三个窗口显示。

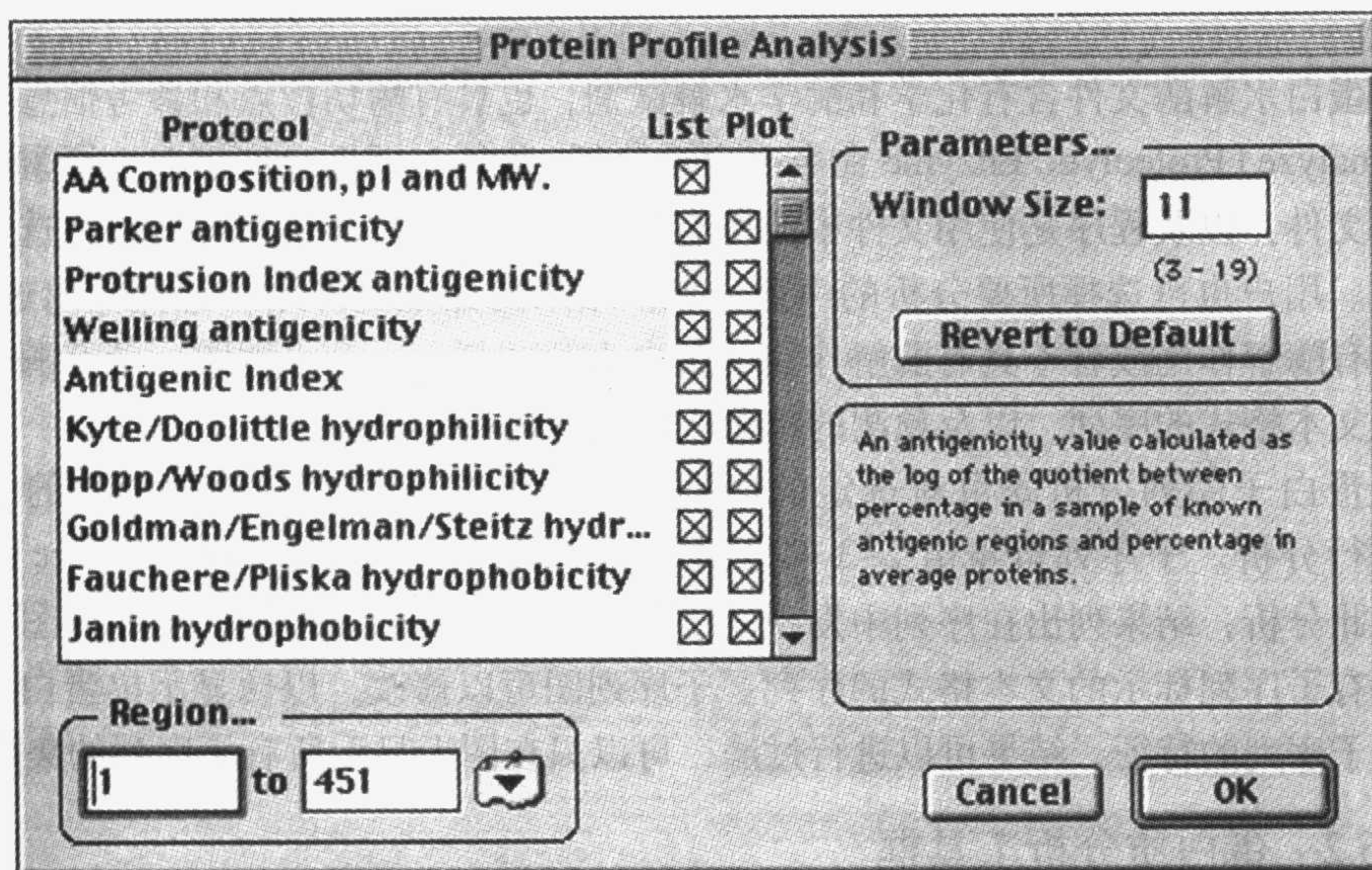


图 4.6 蛋白质分析工具盒对话框

当它的名称在第一个面板中显示时，一个简要的算法说明出现在第四个面板中

4.3.7 序列比较

序列比较范围较广，既可在两个序列之间进行相似检测，也可在一单个序列中甚至整个数据库中进行。MacVector 软件提供 4 种不同的序列比较功能：多序列比对 Clustal W，提供网络在 Entrez 中进行 BLAST 搜索，Pustell 矩阵分析，以及 Align to Folder 功能。Clustal W 比较的是两个或多个序列的相似性，Align to Folder 比对折叠子中的序列与提交序列的，BLAST 则是在 NCBI 中搜索与提交序列相似的序列，Pustell 矩阵分析是搜索两个给定序列的相似性，以图的形式显示。

4.3.7.1 利用 Clustal W 进行多序列比较

Clustal W[7]比较的是多个序列，并将相同的区域和相似的残基排成一行。这种类型的比对在比较不同来源的序列上是有利的，比如：研究不同种属来源的同类序列之间的差异和相似，结果显示不同种属之间的相同区域都是保守的。它首先对每个序列作配对比较，然后比较多个序列中最相似的一对序列，最后根据它们与那对相似性最高序列的相似程度一个一个将序列加入。

当同类型的序列至少有 2 个打开后，Analyze | Clustal W Alignment 命令才可以使用，此命令对话框分 4 个窗格：Pairwise Alignment、Multiple Alignment、Sequences to Align 和 Protein Gap Parameters(图 4.7)。Pairwise Alignment 面板里的参数是控制最初比较的速度(也是敏感度)，这些参数是：评价基数(在

BLOSUM 30，PAM 350，MD 350 或 Identity 之间选择)、队列速度(慢或快)，以及在队列中插入和延伸缺口的补偿。Multiple Alignment 参数是控制最后多序列进行队列时的速度，包括评价基数、开放和延伸缺口补偿和发散延迟(若某序列相似性低于这参数值，那么此序列的比对就会延迟)。Transitions 参数只在核酸的多序列比较时才显示，且在 Weighted 和 Unweighted 之间穿插(A⇔G 和 T⇔C 之间的转换比 A⇔T、A⇔C、G⇔T 和 G⇔C 之间的颠换有利得多)。Sequences to Align 面板有一卷动菜单，它列出了所有本地机上打开的同类型的序列，当这些序列的所有集合进行比较时，它们才能通过点击文件名被选中。Protein Gap Parameters 只在比较蛋白质序列时才可用，它的参数有缺口分隔距离(两个缺口之间所允许的最小距离)，末端缺口间隔(如果此功能可用的话，在设计缺口间隔距离参数时，末端缺口与内部缺口等同视之)，残基特异补偿(若此功能可用，那么在比对的序列队列中，氨基酸特异补偿就可以在每个位置上增加或减少缺口开放补偿)，亲水性补偿(若此参数可用，这些补偿就有可能增加亲水区内的开放缺口：亲水残基由用户在 Hydrophilic Residues 框中限定)。

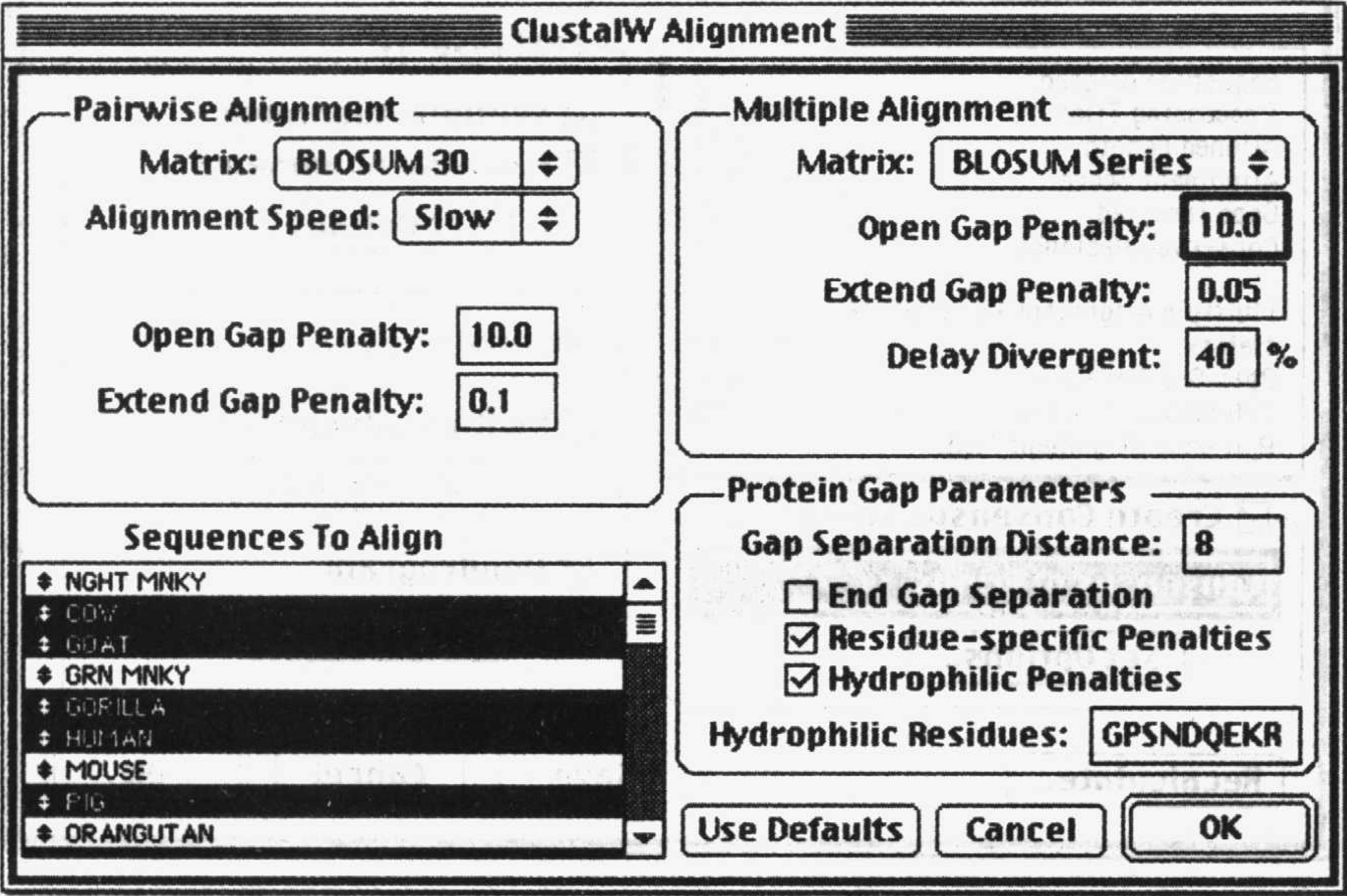


图 4.7　Clustal W 多序列对比对话框

多序列比对编辑器是和结果对话框同时显示的，它允许用户审查和修改比对结果。在比对序列队列中，单个残基的颜色是彩色的，用户可以改变颜色。MacVector 为用户提供了许多种颜色，从化学键到侧链的刚性堆积都有不同的颜色，用户可以选择 MacVector 提供的颜色组或创建新的颜色组。在序列比较编辑器中，参考序列被固定住，以使它们能稳定在最上行，而其他序列则可以在下面

卷动，序列队列本身可以通过增加或移走间隔来编辑，也可以移动序列来改变各个序列的顺序。

结果对话框列出了从序列队列中的各种统计结果，比如所比较序列的数目，序列比较过程所花的时间，序列队列的长度，插入的间隔等(图 4.8)。序列队列可以文本或 PICT 格式来观看，有两个文本格式的结果窗口：多序列比较窗口以 FASTA 格式显示序列队列结果，而配对比对窗口显示的是所有的与序列结合的配对。用户可以选择记号来代表队列中保守和不相同的残基。多序列队列结果以 PICT 格式输出，这种方式的文件质量是公认的，也能修改，相同或相似的残基可用粗体或斜体显示，也可以加框或加上阴影，用户还可以修改残基字母的字体、形式和标题。此外，多序列队列也能以系统进化图显示，序列按树状形式排列，且还能显示它们之间的关系。系统进化图既可以以进化分支图形式显示，也可以以表观分支图形式显示，它们的外形可以被格式化。多序列队列中的一致序列可以保存为 MacVector 序列文件，通过检查框 Create Consensus Sequence 命令即可。

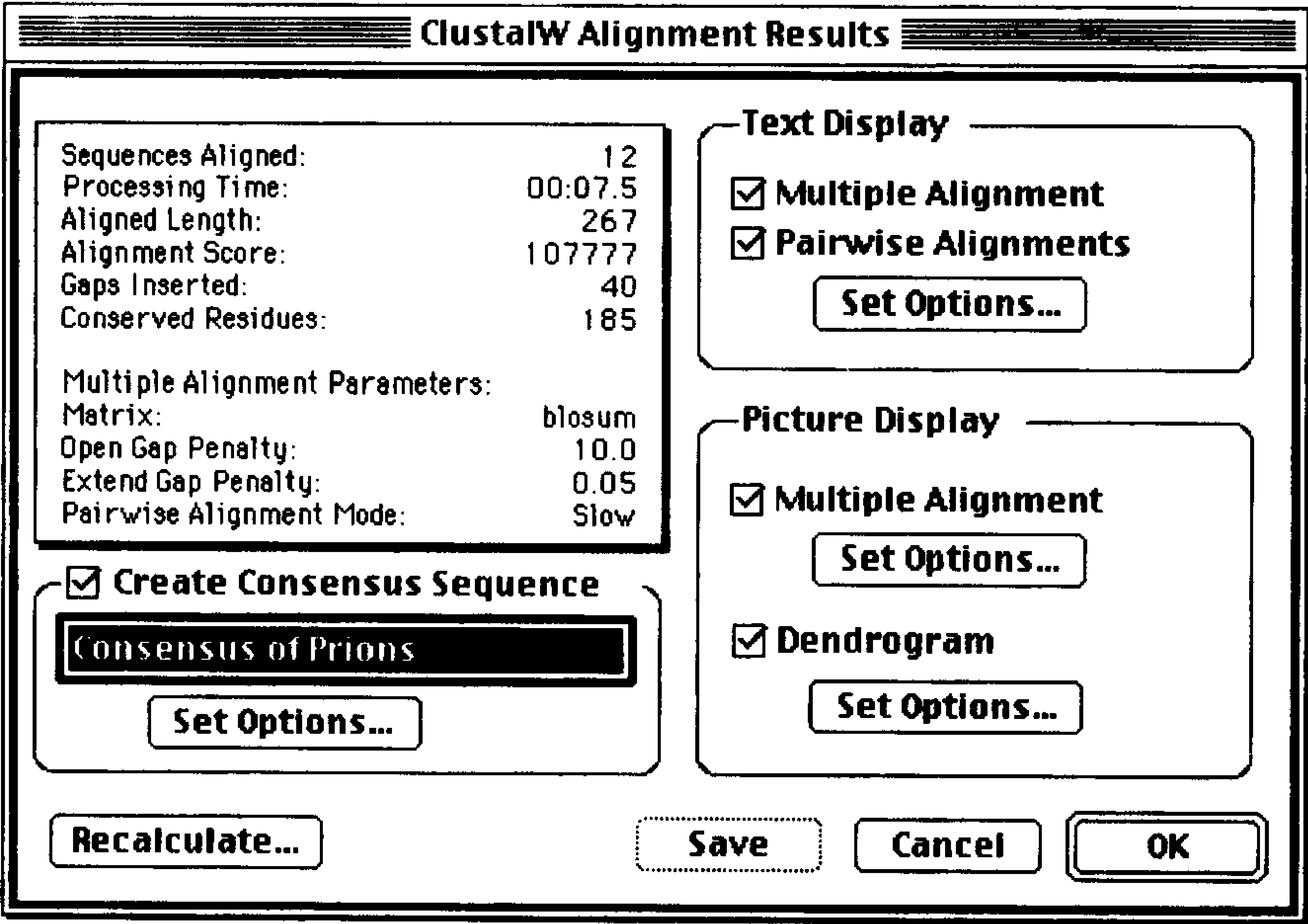


图 4.8 Clustal W 多序列对比结果对话框

4.3.7.2 BLAST 搜索

BLAST[8]搜索可以用来搜索与提交序列相似的存在于NCBI数据库中的序列。MacVector 增加了网络 BLAST 功能，它有以下用途：序列被提交到 NCBI 中的 BLAST 服务器上，并在此进行搜索，结果通过网络返回到 MacVector 上。有 5 种 BLAST 程序可以用来寻找提交序列的同源性：blastn(比较提交的核酸序列与数据

库中的核酸序列)，blastp(比较提交的氨基酸序列与数据库中的氨基酸序列)，blastx(比较用 6 种读框翻译所提交的核酸序列的产物与蛋白质数据库中的序列)，tblastn(比较提交的蛋白序列与用 6 种读框翻译数据库中的核酸序列而成的蛋白质序列)，tblastx(比较用 6 种读框翻译过来的核酸提交序列与用同样方式翻译过来的核酸序列数据库)。

启动 Database | Internet BLAST Search 命令就可执行搜索，若计算机已经连接到网络，那么 BLAST 搜索的对话框就会出现，且伴有程序数据库和 Expect 值选择，数据库的选择是建立在程序的选择上(blastn、blastx 或 tblastx 比较提交的核酸序列，blastp、tblastn 比较提交的氨基酸序列)，以及依赖于分析时 NCBI 中可利用的数据库。Expect 值是一具有统计意义的衡量标准，是报告数据库中不配对序列的门槛，只有配对数达到 Expect 值才能在数据中找寻到。

当搜索完成时，结果对话框列出其统计结果，包括在数据库中搜索到的序列数，观察、保存并比较的配对序列数。结果可以观察到配对序列的目录，以及提交序列和比较序列相互配对的队列。配对队列序列可以从数据库中获得，使之变亮，然后打开命令 Datebase | Retrieve to Desktop，一收到此命令，MacVector 就打开到 NCBI Entrez 服务器上的连接，单独的且带有选定的序列的窗口就出现在本地机上，Datebase | Retrieve to Disk 命令将其保存到硬盘上。

4.3.7.3 Pustell 矩阵分析

矩阵分析是将一序列残基沿 X 轴排列，另一序列沿 Y 轴排列，形成一平面图，两序列残基相同时，就在坐标上打一点。矩阵分析是两序列相似性最好的显示方法，因为它能发现其他方法所忽视的序列特性，通过坐标形式显示结果，两个被比较序列即使是最微弱的相似性也可以让人们用肉眼观察到。由于序列是总体上的比较，那些成双的和重新排列的区域相似性很微弱而往往被计算机队列方式所忽视，但矩阵分析可以显示出来。

在 MacVector 中，Pustell 矩阵分析[9]可以用于比较 DNA 和 DNA 序列，蛋白质和蛋白质序列或 DNA 和蛋白质序列。选择 Analyze | Pustell DNA Matrix、Analyze | Pustell Protein Matrix、Analyze | Pustell Protein &DNA 命令就可以进行上述分析。有些参数用来规范这些比较方法：scoring matrix、hash size、window size、minimum percent score 和 jump parameter。scoring matrix 列出的是不同类型配对的权重(DNA Identity Matrix 由 MacVector 提供，用于 DNA-DNA 比较，而蛋白质 Identity Matrix 也由 MacVector 提供，既可用于蛋白质比较，也可用于蛋白质和 DNA 之间的比较)，它可以进行修改，比如某些不匹配的可视为部分匹配，依据推测，如亲水性、电荷和结构，某些不匹配则可视为完全不匹配。hash size(也认为是 k-tuple size 或 word size)决定分析的灵敏度和速度，它能衡量两序列精确配对的长度，且在 MacVector 软件准备对配对区域进行评价和比较之前就已经完成。

jump parameter 与 hash size 紧密相连，当 jump 设置为 1 时，hash 就代表一行中配对完好的残基数；若 jump 为 3，hash 就代表一行中第一个残基配对完好的三联体数。若在起始 hashing 这一步骤中找到了配对序列，接着就是第二步骤，即 scoring 步骤，scoring matrix、window size 和 minimum percent score 这些参数就被利用起来。当 MacVector 找到一个符合 hash 和 jump 设置的匹配，就开始检查序列队列中匹配的那部分序列，通过 window size 来检测这一匹配部分的大小，用 scoring matrix 的数值对其进行评价，若配对百分比数值大于或等于 minimum percent score 值，这一匹配序列就被保存下来。

结果同时以矩阵图和序列队列两种方式显示，矩阵图是比较结果的点阵图，用户可以通过拖动光标放大目的区域来观察区域内放大的点。序列队列陈列显示的是满足或超过 minimum percent score 参数的 *X* 轴与 *Y* 轴序列的比较结果。

4.3.7.4 文件夹搜索

MacVector 提供 Align to Folder 功能，也指单个数据库搜索。通过 Database/Align to Folder 命令进入，将提交的序列与包含一个或多个序列的文件夹进行比较，使用 Lipman-Pearson DNA 比对法或 Wilbur-Lipman 蛋白质比对法[10~12]，这些方法都采用 3 个步骤：第一步为 hashing，将文件夹中的序列与提交序列进行比对筛选，接着是 scoring 和 alignment。在 alignment 步骤中有 3 个参数，打开 scoring matrix 可以进行修改：cut-off score、deletion penalty(缺口开放补偿)和 gap-penalty(缺口延伸补偿)，只有序列最初的值大于或等于 cut-off score 时，才被比较。alignment 步骤中插入缺口时为了提高最后(最优)的评分值，deletion 和 gap penalty 数值从序列队列中的评分值扣除，这样假若缺口可以提高序列队列值，那缺口就可以插入序列中。processing 参数决定队列比较是在搜索开始时还是最后执行。

结果有保存的匹配列表、比对的水平图和配对的序列。匹配列表中的信息包括序列本地名和与提交序列进行比对的原始值和优化值。水平图以图的形式表示比对结果，表明文件夹序列的哪些与提交序列配对。与其他图表结果一样，用户可以利用光标拖动目的区域来放大，比较序列窗口显示的是提交序列和匹配序列之间的优化比对。

4.3.8 浏览 Entrez

NCBI 中的 Entrez 数据库，内有核酸序列、蛋白质序列和相关的文献信息，序列资料包括来自 GenBank、EMBL、DDJB、PIR、PRF、PDB、SWISS-PROT、dbEST 和 dbSTS 数据库，以及美国、欧洲专利库的全部核酸和蛋白质序列资料。Entrez 还有一个 MEDLINE 数据库集合：序列数据库引用的文献和摘要以及其他相关 MEDLINE 记录。MacVector 通过网络可以浏览 Entrez，Entrez 浏览器具有便

利的特性，因为它允许用户以 MacVector 格式从 Entrez 数据库中直接摘录序列至屏幕上。

通过 Database | Browse Entrez | via Internet 命令就可进入 Entrez 浏览，数据库中有 DNA、蛋白质和 Medline 的识别符号，首先它们中的一个被选定，顶端的 3 个下拉菜单允许用户制定一个或复合注解进行搜索，第一和第三个下拉菜单列出所有可能的搜索种类。对于单个注解搜索，用户可从第一个下拉菜单选择一条目录，然后在启动搜索前于目录菜单下的框中输入搜索行，而进行复合注解搜索，用户从第一和第三个下拉菜单选择搜索目录，从第二个菜单中选择 and 或 not(图 4.9)，一旦搜索行被回车，搜索就开始了。

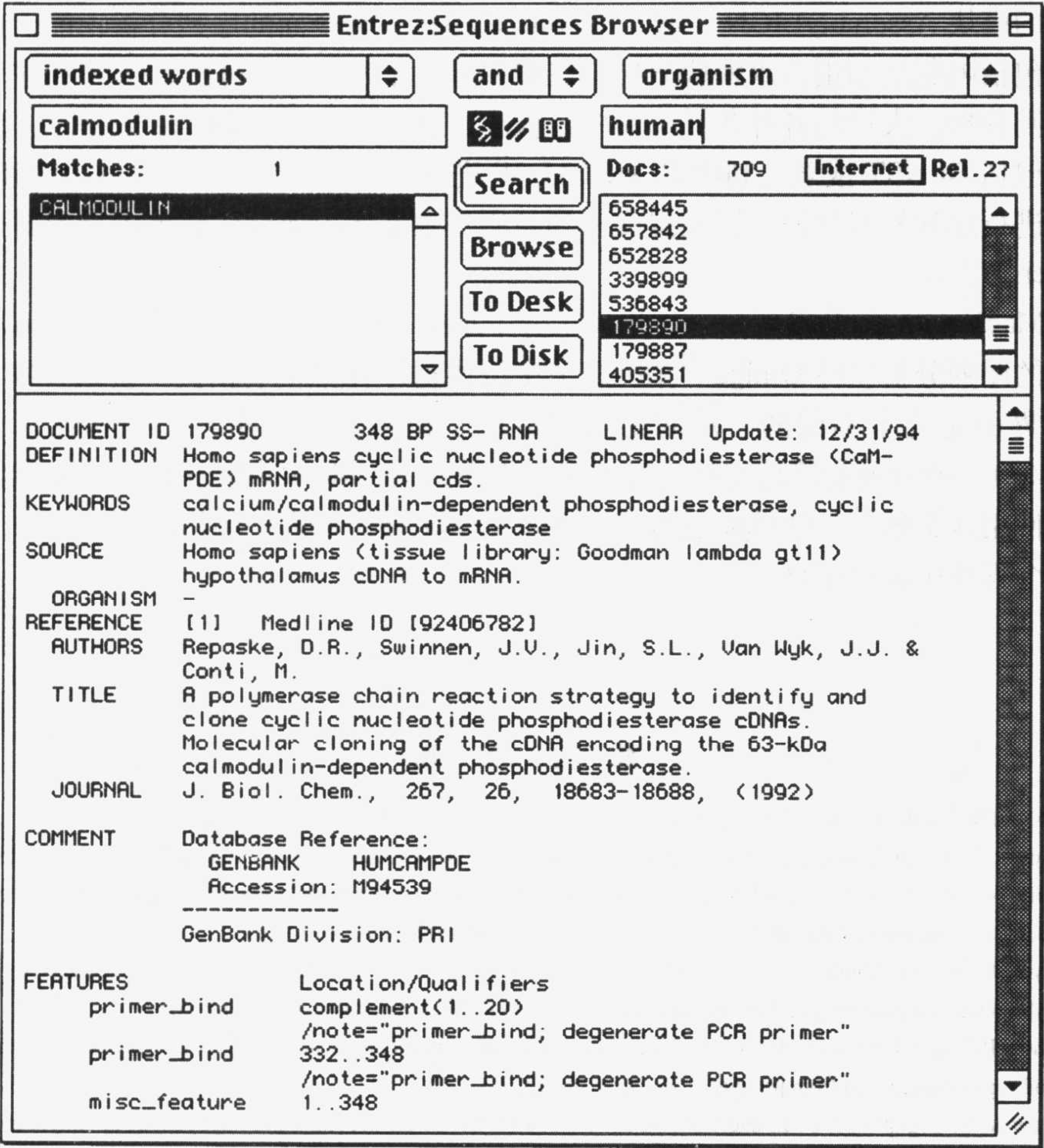


图 4.9 MacVector 中 Entrez 浏览窗口

若有与搜索行匹配的任何结果，都会在两个窗格中列出：左边 Matches 窗格，

包含一个与搜索行相符的目录，只要使用其中任何一个目录就可激活它；右边的 Document ID 格，包含一个序列的 ID 目录，对应于左边窗格激活的目录，按住 shift 键，可使多个相符的目录或 ID 选中并变亮，点击 Browse 按钮，就在窗口的下半部出现一个大窗格，列出了所选并变亮的序列的注解及特性信息，选定的序列可摘取至台式机上，成为一个单独的序列文件。完成此步只需点击 To Desk 按钮，而点击 To Disk 按钮，它就可以移至硬盘下一个已经存在或新建的文件夹中。

4.4 注

MacVector 6.5 是一个在 Macintosh 上编写复杂但易于操作的序列分析软件包，随它的模块 AssemblyLIGN 一起，为 Mac 机提供一套全面的序列分析功能，这些功能包括限制酶和蛋白水解酶分析，核酸和蛋白序列的基序分析，PCR 和序列引物预测分析，大量的各种各样的蛋白分析法，通过网络浏览 Entrez 数据库，序列比较包括 BLAST 搜索，利用 CLUSTAL W 进行多序列分析，使用 Pustell 矩阵点阵法进行序列配对比较，以及提交序列与用户数据库中保存的数据进行配对比较 (folder 搜索)。

MacVector 的开发者——牛津大学分子小组，也编写了一个用于 Windows™ 系统的序列分析软件 Omiga™(在 4.3 节中讲述到)。虽然两者由同一公司开发，但用户界面是完全不同的，从而给用户造成了在两者中切换的一些困难，然而 MacVector 的开发者坚持选择这种苹果机界面，以便 Macintosh 平台用户能非常容易地使用这个软件。Omiga 也可以以 MacVector 格式输入、输出序列，允许用户在两个程序中流通信息。

(廖晓萍　译)

参 考 文 献

[1] Oxford Molecular Group (1998) *MacVector 6.5 User Guide.* Oxford Molecular, Oxford, England.

[2] Fickett, J. W. (1982) Recognition of protein coding regions in DNA sequences. *Nucl. Acids Res.* **10,** 5303-5318.

[3] Gribskov, M., Devereux, J., and Burgess, R. R. (1984) The codon preference plot graphic analysis of protein coding sequences and prediction of gene expression. *Nucleic Acids Res.* **12,** 539-549.

[4] Rychlik, W. and Rhoads, R. E. (1989) A computer program for choosing optimal oligonucleotides for filter hybridization, sequencing and in vitro amplification of DNA. *Nucl. Acids Res.* **17,** 8543-8551.

[5] Chou, P. Y. and Fasman, G. D. (1974) Conformational parameters for amino acids in helical, beta-sheet, and random coil regions calculated from proteins. *Biochemistry* **13,** 211-222.

[6] Garnier, J., Osguthorpe, D. J., and Robson, B. (1978) Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. *J. Mol. Biol.* **120,** 97-120.

[7] Thompson, J. D., Higgins, D. G., and Gibson, T. J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22,** 4673-4680.

[8] Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1994) Issues in searching molecular sequence databases. *Nat. Genet.* **6,** 119-129.

[9] Pustell, J. and Kafatos, F. C. (1982) A high speed, high capacity homology matrix zooming through SV40 and polyoma. *Nucleic Acids Res.* **10,** 4765-4782.

[10] Lipman, D. J. and Pearson, W. R. (1985) Rapid and sensitive protein similarity searches. *Science* **227,** 1435-1440.

[11] Pearson, W. R. (1990) Rapid and sensitive sequence comparison with FASTP and FASTA. *Meth. Enzymol.* (ed. Doolittle, R. F.) **183,**63-98.

[12] Wilbur, W. J. and Lipman, D. J. (1983) Rapid similarity searches of nucleic acid and protein databanks. *Proc. Natl. Acad. Sci. USA* **80,** 726-730.

5 DNASTAR 的 Lasergene 序列分析软件

Timothy G. Burland

5.1 引言

Lasergene 由 8 个应用程序组成，每个程序都组织成功能单元。整个 Lasergene 系统包括以下一些软件：

SeqManII: 整理(trim)和组装(assemble)序列数据，以及确定共有序列。

GeneQuest: 在小的、与 BAC 相当大小的或较大的序列中发现并标注基因、图谱(pattern)以及其他特征。

Protean: 预测蛋白质二级结构，鉴定抗原(antigenic)区。

MegAlign: 以配对形式(pairwise)或多重比对形式分析，并构建系统进化树。

GeneMan: 利用由序列相似性、共有序列和文本术语(text term)构成的布尔查询(Boolean queries)搜索序列数据。

PrimerSelect: 设计用于 PCR、测序、杂交及转录的引物。

MapDraw: 构建限制图谱用来显示位点、翻译(translation)和特征。

EditSeq: 从其他应用程序引入序列并用于 Lasergene 的分析。

以下是使用 Lasergene 应用程序的范例。无论软件运行在 Windows95/98/NT 上还是 Mac 计算机上，其使用步骤是一样的。以 Lasergene99 的 4.0 版为例。

5.2 用 SeqManII 进行序列组装(assembly)

SeqManII 是一个组装软件，与其他软件一起，被用来进行大肠杆菌基因组测序计划[1]。它可以组装小到 1kb，大到细菌基因组的测序结果。本节将举例描述如何把大肠杆菌基因组计划 1200ABI®格式的序列记录文件组装成一个 93kb 的重叠群。

5.2.1 序列登录(entry)

向 SeqManII 项目加入大量序列文件最简单的方法是从 Windows Explorer 或 Macintosh Finder 中拖放。可加入多个文件夹，文件可以是 ABI 和 SCF3 记录文件，

或 DNASTAR 序列文件。SeqManII 最多可接受 64 000 个序列，足够容纳 5 倍的大多数细菌基因组的鸟枪(shotgun)序列。序列加入后，就会生成一个文件名文件(fof)，以便向未来的集合中加入相同系列的序列时，可以加入 fof 而不必加入独立的文件。

5.2.2 载体和宿主数据的整理

序列数据可能会被来自生长克隆的宿主序列污染，也可能被用于克隆靶序列的载体序列污染。这些污染能严重损害序列组装(assembly)。SeqManII 在组装前先除去污染的载体和宿主序列。为从读取的大肠杆菌序列中去除载体，序列被分为两组再进入 SeqManII。当克隆进 Janus 载体[2]时正向读取的序列被标记，反向读取的用载体目录中的 InvJanus 载体标记。在整理序列与载体的相似性时，推荐采用缺省的严谨性。

5.2.3 质量整理

在整个长度上序列读取的质量不同，质量差的数据通常出现在靠近每个读取序列的 5′和 3′处。质量差的区域常含有碱基读取错误，可损害序列组装。SeqManII 直接从荧光记录数据估计最高质量，并清除那些低于指定的质量底限的数据[3]。保留优良数据和清除较差数据都不需要人为编辑。在本大肠杆菌范例中，质量整理严谨性设为推荐的中间值。

整理后，会保存到另一个 fof 中，它保存整理信息和文件名。因此，在包含同一序列的另外集合不再需要重复整理。

5.2.4 序列组装

如果所有读取的序列来源于一个连续的 DNA，而且每个 DNA 片段都产生了读取数据，就可能使这些读取数据组装成一个连续段或重叠群，对应于原始的 DNA 段。序列组装时，可调整很多参数，如加入读取序列到重叠群所需的匹配范围，以及在重叠读取序列的排列上产生断裂的惩罚。对大多数数据，缺省值可产生最好的组装——它们产生最少数量的重叠群，并使错误连接的可能性降至最低。

选择了可调整的参数后，单击鼠标，SeqManII 可完成载体和质量整理并进行序列组装。在 200MHZ 的 PentiumPro™ Windows PC 或 Macintosh G3 上，SeqManII 在 15~20min 内，就可整理和组装 1200 个大肠杆菌读取序列，形成一个 93kb 的重叠群(图 5.1)。

5.2.5 获取共有序列(consensus calling)

序列比对组装后(图 5.1)，将获取共有序列。SeqManII 与其他程序一样，能通过选择给定位置出现最频繁的碱基确定共有序列，即基于大多数的获取体系

(majority calling system)。但如在本项目中，如可获得荧光记录数据时，SeqManII通过估计峰的质量确定共有序列。这个体系比基于大多数的共有序列获取体系更准确。不同于基于大多数的体系，SeqManII在大多数最初获取的碱基发生错误时也可正确地确定共有序列(图 5.1)。人类基因组计划要求 99.99%的精确性，上述功能对适用于这个目标的组装软件是必不可少的。

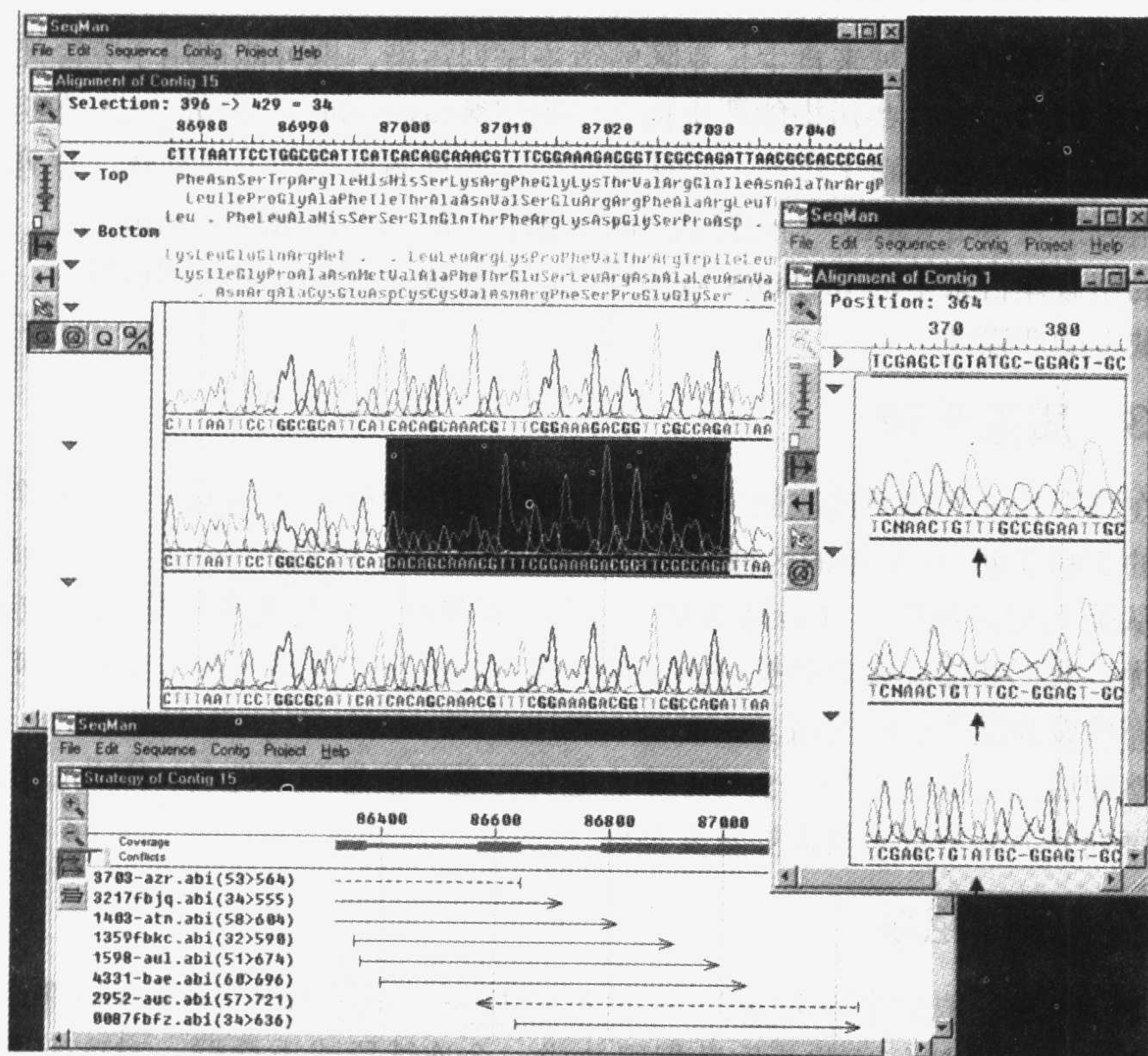


图 5.1　左边显示了 93.8kb SeqManII 项目。左上是比对窗口，包括记录(在彩色显示器上为 4 色)和共有序列及其 6 个翻译框架，翻译中的点为终止密码子。左下是策略视窗，表示每个序列在何处与重叠群重合。右侧的比对视窗显示了一个小项目，基于质量的共有序列获取在共有序列中正确地确定了一个 A，而在对应的栏中获取的碱基为 T、T、A(箭头)。基于大多数的共有序列获取将把此位置错误地确定为 T

不必重复组装过程就可改变共有序列获取的标准。对记录数据，利用证据百分率参数(evidence percentage)控制共有序列获取的严谨性——即特定的碱基要求(call)一个国际生物化学协会(IUB)模糊码代表一个碱基差异的或可疑的位置。把证据百分数设高，将提高碱基差异或模糊的可能性。设定过高可导致虚假的碱基差异或模糊。相反，过分降低证据百分数，当获取碱基的证据模棱两可或位置出现碱基差异时，将提高获取确定共有序列的风险。

5.2.6 编辑

尽管数据整理、组装和共有性确定可自动进行。SeqManII 提供一个图形界面，手工编辑读取的序列和共有序列(图 5.1)。读取的序列、记录和 6 个翻译框都与共有序列排列在一起。翻译中可读框和终止密码子指示了潜在的框架移位或其他序列问题。

策略视窗(图 5.1)显示在重叠群中序列重叠的范围和方向。用户可选择完全和部分重叠的阈值。在本大肠杆菌项目中，完全重叠定义为四重重叠，在每个方向上至少有 2 个读取序列。重叠群上不同的颜色和粗细显示在何处只是一个读取序列或只在一条链上重叠。在本例中，在 87 000 核苷酸周围为完全重叠，但附近的区域则相反(图 5.1)。策略视窗简化了以下决定，即是否需要另外的实验以完成项目。

5.2.7 进一步的分析

对重叠群满意后，可选择任一或全部序列通过互联网利用 NCBI 的 BLAST 服务器进行 BLAST 查询。返回结果为与重叠群相符的详细公共序列。这是序列性质有用的初步指示，如果序列的片段与公共数据紧密相关，可作为评估组装效果的一个独立方式。

重叠群可按 DNASTAR、GenBank 或 FASTA 格式输出文件。对于在全长不能与公共数据紧密相配的序列，下一步需要用 GeneQuest 发现基因。

5.3 利用 GeneQuest 发现基因

GeneQuest 识别 DNA 序列广泛的特征，并提供注释和可视化工具。鉴定特征后，信息提交公共数据库。

5.3.1 序列输入

GeneQuest 接受 DNASTAR、GenBank、ABI 和 SCF3 格式的序列文件。项目可以是任何大小，大到整个细菌基因组。本节举例说明如何从 GenBank 中的一个 28.67kb 的秀丽新小杆线虫(*Caenorhabditis elegans*)的黏粒克隆(登录号 no. Z46240)中寻找基因。

5.3.2 发现编码区

5.3.2.1 重复序列和碱基分配

高重复区域不太可能编码蛋白质，因此发现重复序列可在搜寻基因时排除这

些片段。GeneQuest 识别正向、二元(dyad)和反向重复序列。在秀丽新小杆线虫的 20~21.7kb 存在同向重复，表明在此区域没有编码潜能(图 5.2)。

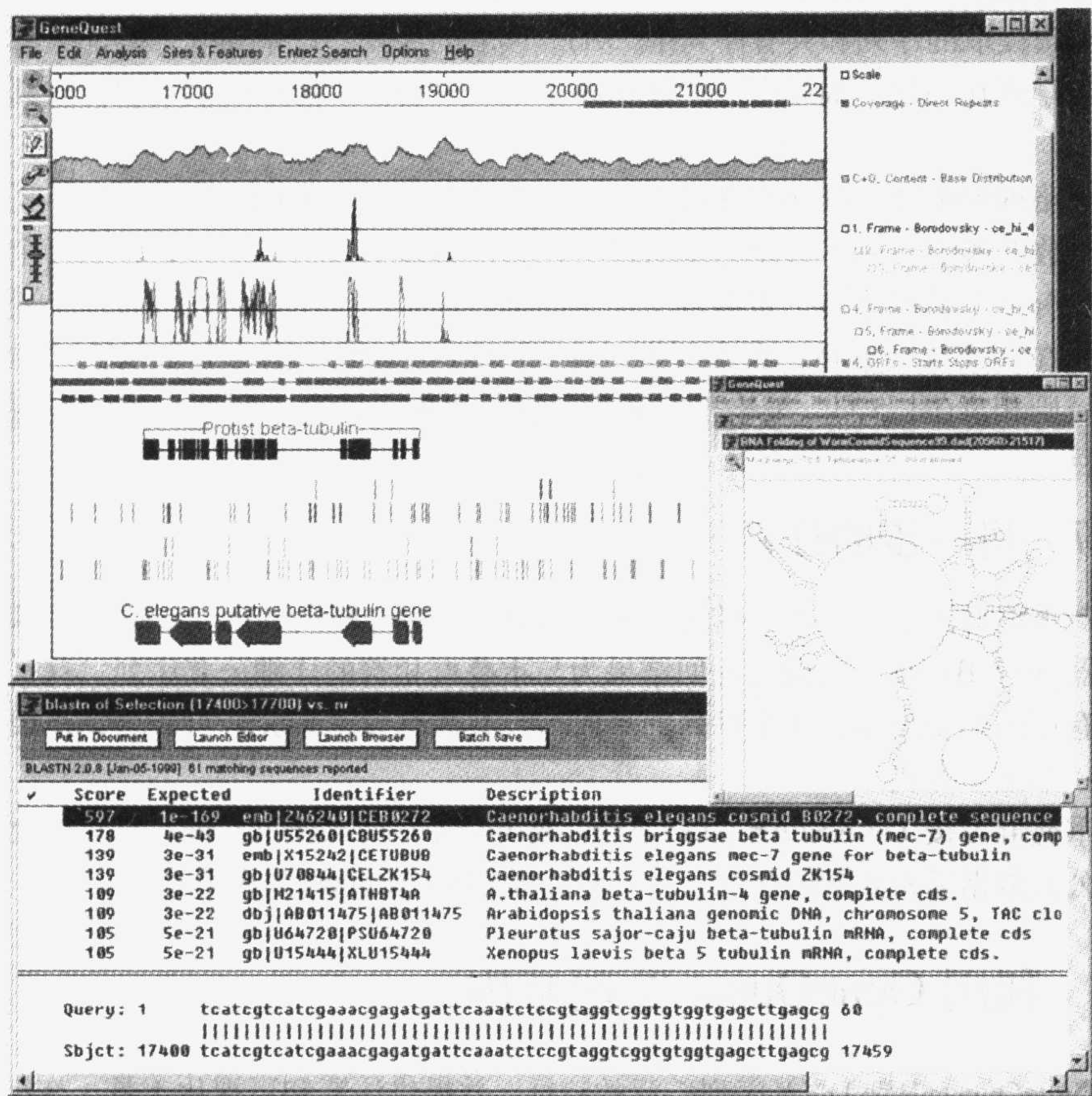


图 5.2　上方窗口是 GeneQuest 显示方法结果的主视窗——“分析表面”。上方标尺下第一行显示正向重复。其下方是 G+C 含量图。再下一个图是 Borodovsky 统计，为框 1~3 的叠加，没有编码潜能的证据。其下的条棒显示了框 4~6 的可读框，其正上方是同一框架的 Borodovsky 数据，与存在多外显子基因一致。下面的 Gene Finder 显示原生生物β微管蛋白符合 Borodovsky 峰。通过观察剪切位点前后的数据(另一些杂乱标记)，我们可以预测内含子和外显子的边界。提供序列的作者在 GeneQuest 发现基因的相同位置注释了一个微管蛋白基因。右下显示了预测的重复序列的 RNA 折叠，左下显示了利用跨越 17 700~17 400 的可读框进行 BLAST 查询的结果

在编码区和非编码区 G+C 含量不同的生物中，确定碱基含量有助于基因鉴别。本例中，50 核苷酸的窗口用于计算平均 G+C 含量(图 5.2)。在整个序列上 G + C 含量变化并不是随机的，高 G+C 含量的区域与低 G+C 含量的区域交替存在，可首先在高 G+C 含量的区域上发现基因。

5.3.2.2 可读框(ORF)、终止、起始

原核生物缺乏内含子，可通过寻找起始密码子和终止密码子间的最长的可读框的方法从序列数据中鉴定编码区域，尽管这个简单的方法不足以作为发现基因的工具。真核生物内含子是常见的，外显子只占构成基因的序列的一小部分，甚至真正的可读框也常常比基因小得多。图 5.2 显示了秀丽新小杆线虫的可读框和终止。GeneQuest 显示了大量的候选编码区，但很少可读框长度超过 500 核苷酸。

5.3.2.3 编码区——Borodovsky 统计

鉴定候选编码区的统计学方法功能很强，对发现真核生物编码区是必要的。Borodovsky 方法[4]发现编码区的序列模式特征，本方法使用秀丽新小杆线虫专一性的矩阵(matrix)文件生成 6 个读框的 Borodovsky 图形(图 5.2)。统计学强烈地显示了对应于序列 19 000~16 500 的框架 4、5、6 的编码能力(图 5.2)。只有极少量证据表明了框架 1~3 的编码能力，而在这个区域两侧 2 kb 完全没有证据。这些结果符合框架 4、5、6 上存在多外显子基因的事实。

5.3.2.4 相关发表的序列——BLAST 搜索

为了尽快确定序列可能编码什么蛋白质，可选择具有编码潜能的区域，如秀丽新小杆线虫核苷酸 17 700~17 400 片段，作为查询通过互联网进行 GenBank 的 blastn 搜索。本例中发现序列十分匹配，因为它已经发表了(图 5.2)。所有与查询十分匹配的都是β微管蛋白，且匹配十分明显。这表明用来查询的片段可能是β微管蛋白基因的一部分。

5.3.2.5 寻找特殊的基因——Gene Finder

本例中预测的编码区涉及已知序列，可使用 GeneQuest 的 Gene Finder 功能明确地检验其与β微管蛋白的亲缘关系。选择原生生物 prototypical β微管蛋白多肽文件(GenBank 登录号 No. M58521)作为要发现的蛋白质。图 5.2 表明这个微管蛋白在哪些区域与秀丽新小杆线虫编码相配。与成簇的 Borodovsky 峰有明显重叠，强烈暗示β微管蛋白的存在。

5.3.2.6 剪接位点——统计学模式

GeneQuest 提供定位内含子-外显子边界的统计学方法。本例中，用秀丽新小杆线虫专一性的矩阵文件预测潜在的剪接位点(图 5.2)。

编码区域和内含子-外显子边界的精确界限可通过观察可读框前后的预测的剪接位点来研究，理想的情况下，Borodovsky 峰与可读框在同一可读框里一致，以供体和受体剪接位点分界。如果用统计学模式方法不能发现候选的剪接位点，

可以放大 GeneQuest 的显示，以区分单独的碱基，再通过肉眼寻找适于做剪接位点的二核苷酸。

把从这个秀丽新小杆线虫获得的信息放到一起，可把微管蛋白的编码区分为 7 个外显子(图 5.2)。本例中序列已经发表，可检查发表的注释。事实上，提交 GenBank 的作者已在 GeneQuest 预测的同一位置注释了推测的 7-外显子β微管蛋白基因。

5.3.2.7 其他发现和注释功能

除了在秀丽新小杆线虫例子中引用的方法外，GeneQuest 可识别转录因子结合位点、限制位点，用户输入的任何模式、全部或部分序列的密码子使用情况、易弯曲的 DNA 区域[5]。GeneQuest 也可模拟限制片段在琼脂糖凝胶电泳中的分离情况，并预测对应于选择的 DNA 的 RNA 的折叠情况。

5.3.3 通过文本(text)搜索以发现相关序列

GeneQuest 提供对国家生物技术信息中心(NCBI)的 Entrez 服务器的访问，可通过互联网处理序列数据的文本查询并返回结果。例如，为了从秀丽新小杆线虫发现其他的微管蛋白，可建立对“文本区‘tubulin’”和“生物种类区‘新小杆线虫属’”的核苷酸数据 Entrez 查询。

在 1998 年中有 68 条数据库登录，包含秀丽新小杆线虫的α、β、γ微管蛋白，种内其他成员的微管蛋白，微管蛋白酪氨酸连接酶，及与微管蛋白有关的其他登录。但 Entrez 服务器不支持如 GeneMan 提供的像布尔文本/序列搜索那样的复杂查询。

5.3.4 注释和显示特征

GeneQuest 提供显示和注释特征的工具，无论它们是用 GeneQuest 发现的，还是在输入到 GeneQuest 的公共数据文件中的。当用范围选择工具选择感兴趣的区域，当 Features_New Feature 菜单被选择时，GeneQuest 自动向注释窗口输入候选序列。可加入从标准特征描述到自由形式注释的一切信息。在本例中为了注释微管蛋白，选择第一个外显子并作为特征注释，然后选择其他外显子加入到第一个外显子后。如果有关任何性质边界的结论发生变化，在注释窗口候选序列也相应调整，并保存重新建立的注释结果。

根据特征不同，可以不同的形式显示，包括图形、盒子、箭头、棒、杂点、文本(图 5.2)。可从 GeneQuest 工具栏选择颜色和填充模式，通过上下拖动，图形元件可重叠、重排、并列。一个有用的方法是为读框 1、2、3 指定颜色如红、蓝、绿色，然后对框架专一性的元件，如可读框、终止、起始和 Borodovsky 图始终采用一致的颜色。

5.3.5 为再使用而创建一系列方法

在 GeneQuest 中分析方法的范围是很广泛的，而且改变方法参数的范围极大扩展了用于分析 DNA 序列方法的组合。当一系列方法采用特殊参数应用于一个序列，GeneQuest 能保存方法大纲(method outline)，与字处理过程中使用的模板类似。通过应用保存的方法大纲，可很快地对另一个序列采用相同的方法和参数。

5.3.6 进一步分析和结果发表

使用外部插图功能，GeneQuest 图形视窗可以拷贝到剪贴板，然后使用 Paste_Special(选择性粘贴)选项(选择贴图而不是序列)移动到其他应用程序。为了在 Microsoft Powerpoint 中编辑图像，可使用拖动(drawing)工具分解(ungroup)粘贴的图像。每个元件——图、棒、箭头、标签、图例等，可被独立地编辑、重置大小、移动、删除。

GeneQuest 把项目保存为 GeneQuest 文档，为了在 Lasergene 等其他应用程序中进一步分析，数据也可保存为 DNASTAR 或 FASTA 文件。为了提交到公共数据库，序列和注释可保存为 GenBank 单调文件(flatfile)。

5.4 用 Protean 进行蛋白质结构分析

Protean 与 GeneQuest 工作方式相同，只是处理多肽而非 DNA 序列。Protean 接收 DNASTAR 格式的序列文件。对于其他格式的文件，应用 EditSeq 模块将序列转化成 DNASTAR 格式(见 5.7 节)。

5.4.1 分析序列

Protean 有超过 20 种分析方法预测蛋白质的二级结构和物理化学性质。与 GeneQuest 一致(见 5.3 节)，大多数方法提供用户定制的结果图形显示。本节以人的钙调蛋白为例分析。

5.4.1.1 预测α螺旋、β片层、无规卷曲、转角

Protean 提供 4 种方法预测二级结构[6~9]，本例采用 Garnier-Robson 方法[7]预测螺旋和转角。钙调蛋白是熟知的蛋白质，所以可通过 Protean 进行基于计算机的螺旋和转角预测并与实际情况比较——在本例中有良好表现(图 5.3)。对于未充分鉴定的蛋白质，需要采用多种方法分析二级结构。

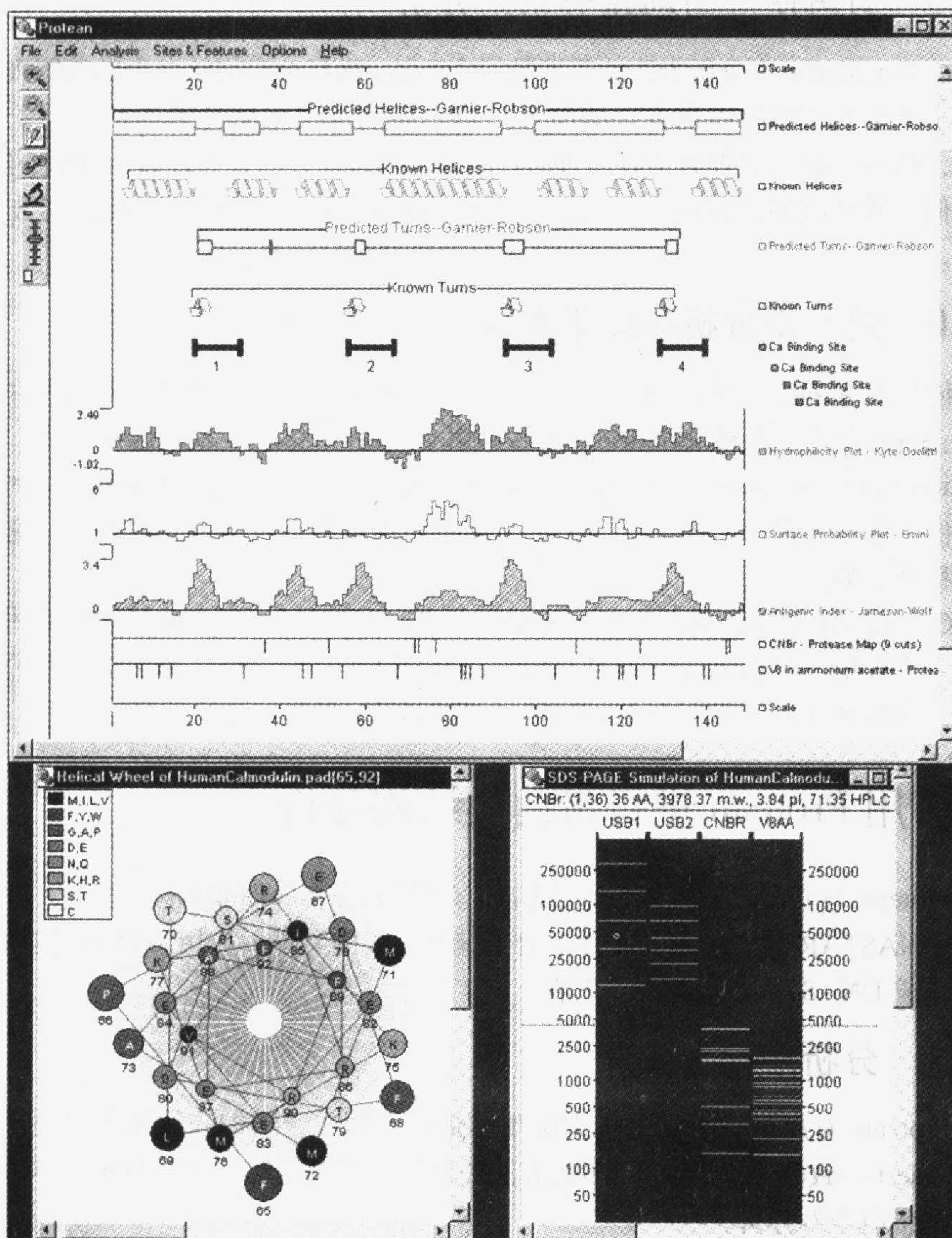


图 5.3　上方的窗口显示了 Protean 的主视窗。右下是被 CNBR 和胰蛋白酶(TRYT)水解产生的序列片段的电泳分离模拟，与已知的分子质量标准比较(USB1、USB2、BRL)；选择带的性质在凝胶图上方显示。左下显示了蛋白残基 65~92 的α螺旋片段的示意图，是假想的观察者从螺旋的中心向下看的情况

5.4.1.2　预测亲水性和两亲性

三种方法用来预测亲水性[10~12]，也可用预测两亲性的 Eisenberg 方法[13]预测

亲水性图谱。在钙调蛋白例子中，通过 KyteDoolittle 方法[11]预测钙调蛋白在其全长都是亲水性的，因此不可能嵌入膜，又一次与真实情况一致。

5.4.1.3 发现基序和序列相似性

通过在 PROSITE 数据库[14]搜索发表的基序是否符合蛋白质中的一个或多个片段，Protean 在钙调蛋白中定位了 4 个已知的“EF-hand”钙结合位点。它们位于预测的转角区域(图 5.3) —— 4 个位点结构相似性的又一证据。

Protean 提供 BLAST 搜索功能。与 GeneQuest 一致，BLAST 搜索的结果在相当大的程度上依赖于选择用全序列还是部分片段作为查询(query)。与基于全序列的查询相比，选择蛋白质序列特殊片段的能力，提高了在无关蛋白质中发现匹配基序或结构元件的概率。

5.4.1.4 预测抗原区域、表面概率和柔韧性(flexibility)

计算机抗原区域鉴定对制备有用的抗体有很大帮助。Protean 提供了 4 种方法预测抗原性[15~18]。Jameson-Wolf 方法[16]在这个小蛋白质中显示了 5 个高度的抗原区域(图 5.3)，其中 4 个与钙结合位点一致。产生的抗体可与其他种类蛋白质的 EF-hand 钙结合位点相互作用。Protean 分析表明若想避免与其他的钙结合蛋白相互作用，在残基 40~50 附近的抗原区是一个很好的抗原选择。

利用免疫细胞化学制造抗体时，表面区域通常是良好的抗原——如果抗原决定簇位于天然蛋白的表面，在抗体产生上会有更好的效果。Emni 方法[19]表明，钙调蛋白残基 40~50 的表面可能性为中等，表明在免疫细胞化学中是一个可接受的目标区。但 Emni 方法也指出残基 75~85 周围的亲水区是一个突出的表面区(图 5.3)。预测此区域有中等的抗原性质，而且并不与任一钙结合位点重叠。相对于 40~50 区域，它可能是更好的抗原。

抗原位点的寻找与内建的 BLAST 查询相结合，能把候选抗原片段限制为具有适合序列专一性的片段。

5.4.1.5 发现蛋白质水解位点

Protean 有 24 个蛋白质水解活性的数据库，包括酶和化学试剂。内建编辑器允许对活性的加入或修改。图 5.3 显示了钙调蛋白被溴化氰(CNBR)和 V8 蛋白酶断裂的位点。与 GeneQuest(5.3.2.7 节)一样，也可显示蛋白水解片段的电泳分离模拟图(图 5.3)。

5.4.1.6 模型结构

Protean 可显示所有或部分序列的模型结构。模型包括 helical wheel、helical net、beta net、linear space fill 及化学式。这些模型提供了关于基因产物三维结构

的线索。

5.4.2 建立重新使用的一些方法

为使采用相同方法组合的重复分析更有效率，与 GeneQuest(5.3.5 节)一样，Protean 可保存并重新使用方法大纲。

5.4.3 结果的发表

与 GeneQuest(5.3.6 节)一样，Protean 的图形视窗可被内建的图形工具改变，或拷贝到剪贴板上供输出到其他应用程序进一步修改。

5.5 用 MegAlign 进行序列比对(alignment)和构建进化树

MegAlign 可比对配对的或多个 DNA 或蛋白质序列，并在比较结果的下方产生序列相似性和差异性的图形显示，以及进化树和数值数据表。

5.5.1 从文件或公共数据库输入数据

MegAlign 直接从 DNASTAR 文件、ABI 和 SCF3 记录文件接受序列。当同时输入 DNA 和蛋白质序列时，MegAlign 把 DNA 序列翻译成蛋白质，并按蛋白质方式排列所有序列，除非设定为把蛋白质翻译成 DNA 进行排列。

如果研究者的蛋白质是新的，把蛋白质序列作为查询传送到 NCBI BLAST 服务器通常返回相似序列的列表。所有这些都可直接加入 MegAlign。

5.5.2 比对多序列

利用 Jotun Hein[23]和 ClustalV[24]运算法则可执行多个比对。在序列比对和产生进化树时，两种方法的结果均略有不同。当必须谨慎地对结果进行解释时，能用两种方法进行结果比较是重要的。

本节例子中，比对一系列细菌的 RecA 序列。用户在 MegAlign 工作表中编辑和排列序列名称，并调整每个序列的片段以利于比对。比对后，工作表提供工具可对比对进行小的调整。两个面板显示比对的左右末端，所以当对比对的一个区域进行调整时，可立即观察对到另一个区域的影响。在工作表中共有序列显示在比对上方(图 5.4)。

一个有用的策略是对一系列完整序列进行比对，根据结果选择共有的亚片段进行最后的比对。图 5.4 显示了 RccA 序列的初步比对。大多数序列区配合得很好，但起始和末尾有很大变化，不适于包含在进化树数据中。因此，序列将用原始或修改的比对参数进行整理和重排。

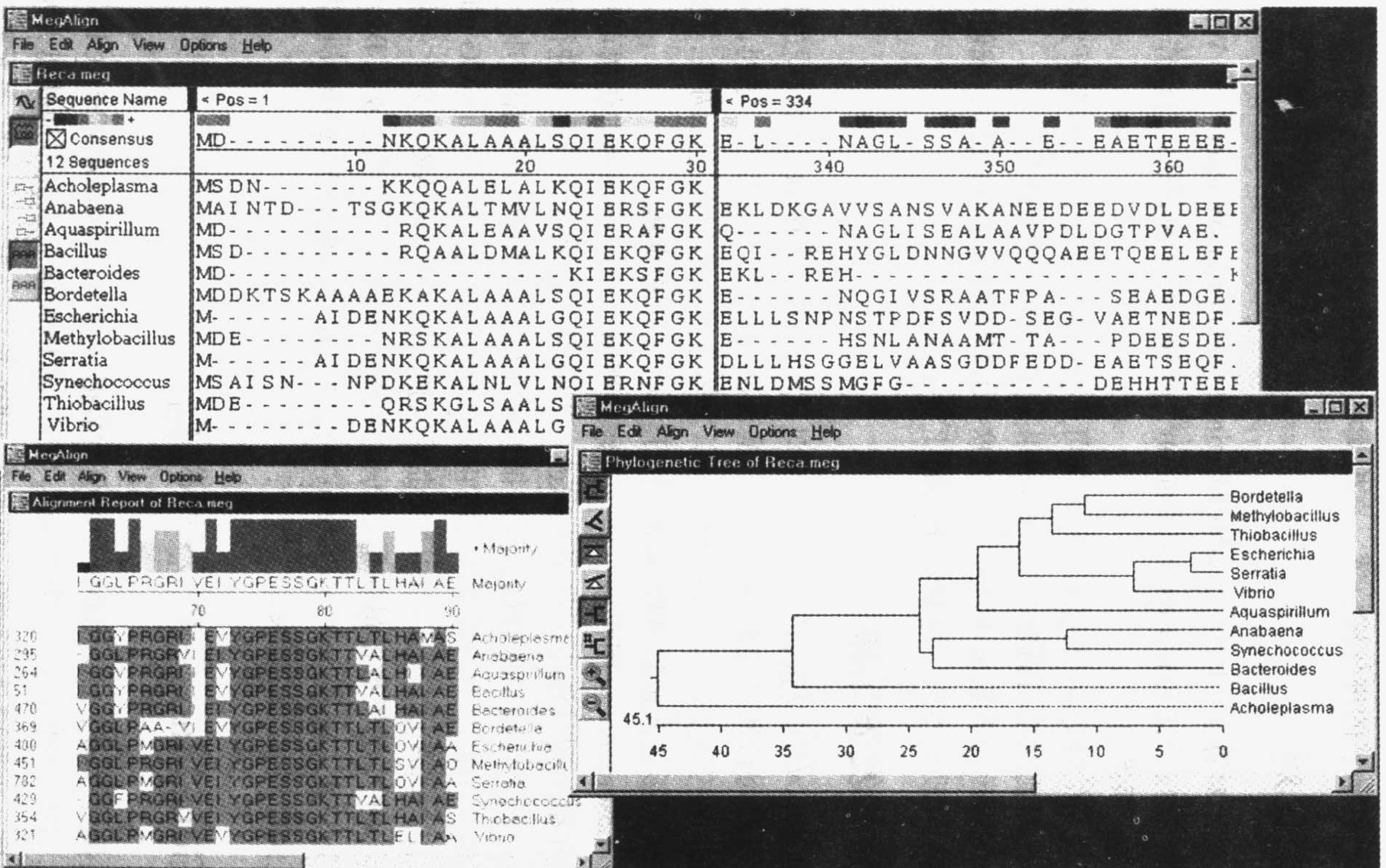


图 5.4　MegAlign 的工作表，在其上可添加和比对序列，也可进行比对的手动调整。两个面板显示了 12 个 RecA 样的蛋白质比对的不同片段。在运行 Windows NT4 的 200MHZ 高能奔腾 PC 上利用 Clustal 算法进行比对需要大约 5s。条状图(bar-chart)显示了每栏共有残基间一致的程度。右下方显示了进化树

用户可使用不同的比对参数重排序列。在比对过程中随时可观察进化树(图 5.4)，这是很重要的，因为对比对的微小编辑和比对参数的变化都对按比对建立的进化树有一定程度的影响。反复改变比对参数并观察产生的进化树可分辨对参数变化不敏感和敏感的进化树位置，在推测种系发生时要更注意后者。

5.5.3 配对比对

有 4 种配对比对算法：DNA 比对有 Martinez Needleman-Wunsch[20]和 Wilbur-Lipman[21]算法；蛋白质比对有 Lipman-Pearson[22]算法；第 4 种算法，打点作图(dot plot)适用于 DNA 和蛋白质。

在工作表视窗的任意两个选择序列间都可进行配对比对，而并不需要去除其他序列。每个配对比对都可调整参数，其结果显示在自身的窗口内。

如果部分配对比对看起来不满意，可对这个片段进行选择和重排，或使用原初的比对参数，或改变参数。这对优化长的比对尤为有用。

5.5.4 观察和发表结果

Alignment Report(比对报告)视窗(图 5.4)对显示在工作表中的相同数据进行用户化的显示。可对每行残基的数目、字体、是否显示序列、和/或共有序列，是否显示比对间相似性的图形表示，以及如何显示个体差异和相似性进行调整，例如，一致的残基可被阴影化或被隐藏以强化不同于共有序列的残基。

与其他的 Lasergene 应用程序一样，工作表视窗、比对报告和进化树可被拷贝到剪贴板上并粘贴到其他应用程序中作进一步阐明。如果需要对数据进行进一步计算分析，序列距离和残基替换表可直接粘贴至微软的 Excel 的数据表中。

为使用其他应用程序中的数据，可将它们存储为适用于 PAUP 和 GCG pileup 程序的格式。在 MegAlign 项目中的序列也可以独立的 DNASTAR 序列文件格式输出。这将是有用的，如果数据最初是从同事的 MegAlign 文件获得，或是 BLAST 搜寻的结果。可分析 GeneQuest 或 Protean 中的独立序列以说明 MegAlign 检查序列中发现的任何序列变化的结果。

5.6 用 GeneMan 发现公共数据

GeneMan 搜寻存储于 CD-ROMs 中的公共数据，包括 GenBank/EMBL，GBTrans 和 PIR/NBRF。用户可搜寻多个 CD-ROMs 目录，或把数据下载到本地硬盘并从单一位置搜寻。后一个选项是可行的，因为在 1998 年可以用大约 400 美元购买到 17G 大小的硬盘。

5.6.1 序列相似性

GeneMan 利用改进的 FASTA 算法，基于蛋白质和 DNA 的序列同源性搜寻公共数据库[25]。GeneMan 从序列文件形成查询公式，并通过选项改变序列定位用于查询，如 k-tuple(重数，比较的单位)、计算同源性的观测窗大小、匹配需要的相似性百分数，以及当查询和数据库匹配时在二者间引入差异的罚分(penalty)(图 5.5)。对于 DNA 查询，在基于 FASTA 的查询前采用快速筛选选项来筛选数据库以降低序列数，随后用 FASTA 正规查询，以加快查询速度。

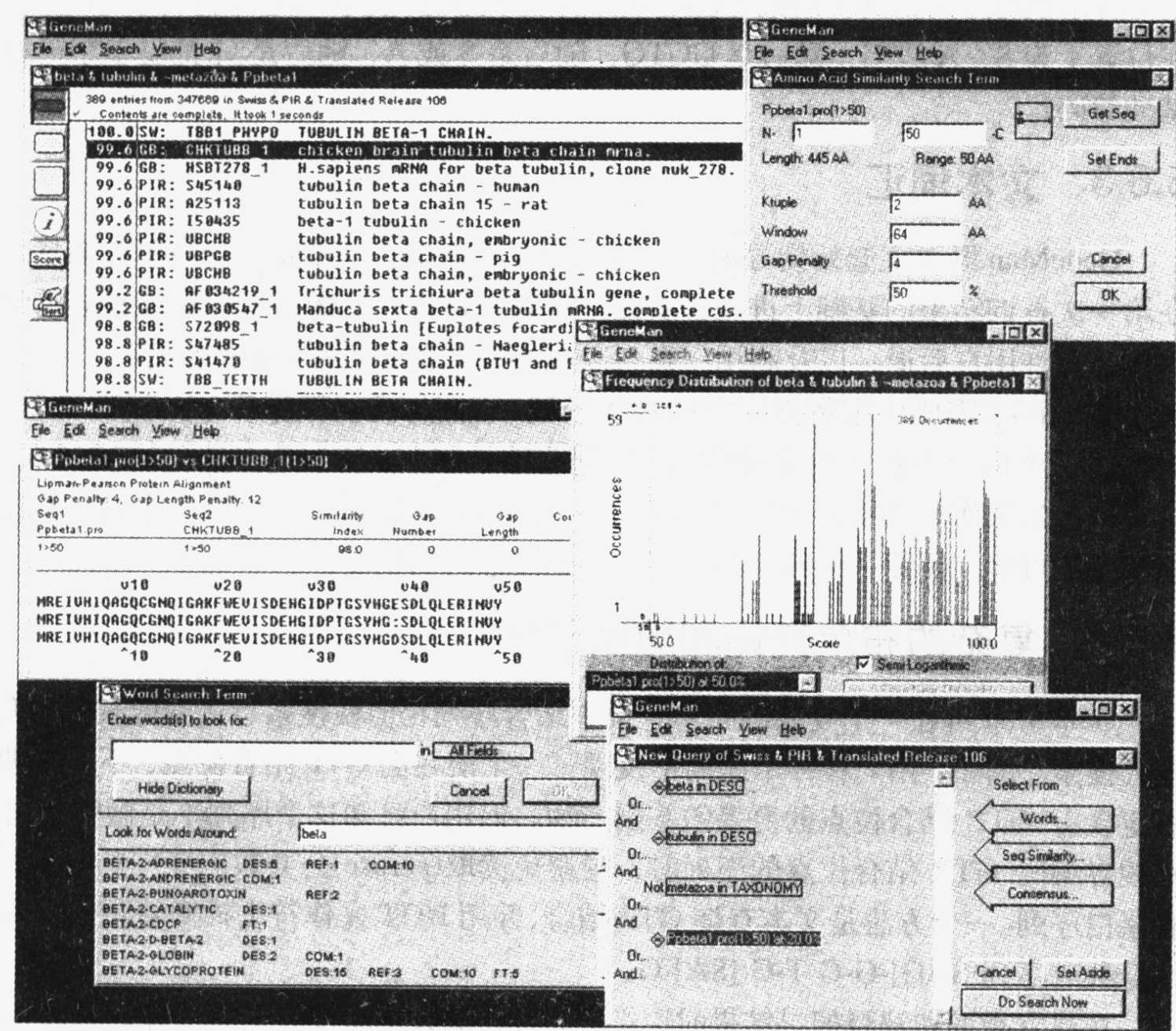


图 5.5　左上是用β微管蛋白查询蛋白数据库序列相似性的部分匹配列表，其下方是最佳匹配比对，可用鼠标选择匹配结果来显示其他任何匹配比对。右上和右下是设置查询和调整序列相似性参数的对话框。左下词搜索对话框和列出索引词的相关字典。右侧中部显示具有不同水平的序列查询相似性的数据库匹配数目的摘要图

最初的查询结果是简单的摘要，用户可选择扩展视图，使之包括更多或全部信息。附加的显示包括数据库符合(hit)频率对相似性百分数的作图，以及查询和

符合的数据之间的比对。

5.6.2 共有序列

GeneMan 的共有序列查询功能支持高达 256 字符的共有序列查询，并允许调整匹配的相似性百分数。共有序列的查询是发现已知基序的有效方法。确定共有序列的语法基于 Prosite 的协议[26]。这些协议支持 IUB 码，对于任何序列位置都允许明确描述可选择的残基、排除的残基、残基间的具体距离以及是否某个模式(pattern)必须位于氨基或羧基末端。

一个共有序列的例子是微管蛋白 GTP 结合共有序列，[SAG]-G-G-T-G-[SA]-G，在位置 1 为 S、A 或 G，随后是 GGTG，随后是 S 或 A，随后是 G(都是单字母氨基酸编码)。

5.6.3 文本词汇

GeneMan 既可在起始搜寻前建立复杂的文本查询，也可先进行简单查询然后在第一次查询结果的基础上进一步查询。后者中每步查询的结果保存在独立的窗口中，允许用户后退，以及利用一系列以前的结果完成多种不同的搜寻。

查询词汇需要确定数据库中的具体领域，否则将搜寻数据库中的所有部分。为确定一个词是否在字典中被索引，可浏览整个字典，从中选择任何词(图 5.5)。可利用布尔操作符 AND、OR、NOT 的组合结合领域限制，以查询多个词。利用这个方法可建立更专一性的查询，提高数据搜索的精确度。

5.6.4 复杂的布尔查询和数据子集

GeneMan 可在一个复杂的布尔查询中结合序列相似性搜索、文本搜索、共有序列搜索。例如，用户在一个序列中发现了一个潜在的 GTP 结合位点，下一步要鉴定含有 GTP 结合位点的更多的基因产物。但用户已知这个序列不是微管蛋白，在搜寻编码 GTP 结合位点的序列时，想避免读取(peruse)公共数据库中的成千微管蛋白序列。一个方法是文本查询 GTP 结合，并用 NOT 操作符去掉微管蛋白 GTP 结合共有序列[SAG]-G-G-T-G-[SA]-G。

在建立查询和解释搜寻结果时应细心，与其他数据库搜寻工具一样，GenMan 不能更正源数据中的错误和遗漏。尽管管理员十分努力，但在许多公共数据库的入口中仍含有错误，而且注释也不完全一致。例如，为发现所有哺乳动物β微管蛋白(β-tubulin)序列，可逻辑上在 Definition field(定义域)确定检索词“tubulin” AND “beta”，而且所有域中确定检索词“哺乳动物”。这种方式可以发现大多数哺乳动物β微管蛋白的入口，但不能发现那些在定义域不含有微管蛋白的β微管蛋白序列入口。但通过搜寻β微管蛋白相似序列，AND 在任何域中检索哺乳动物，可发现哺乳动物β微管蛋白，尽管在定义域中缺少微管蛋白，因为β微管蛋白序列是高

度保守的。

类似的，想像试图查询从符合的列表中去除后生动物(Metazoan)序列，逻辑上这可以通过对文本查询“Metazoa”应用布尔操作符 NOT 限制数据域的来源来完成。但这个查询会产生许多数据库入口的列表，包括一些来源于 Metazoan Drosophila 的序列，这是因为在一些 Drosophila 数据库入口的来源域(source field)中不包含术语“Metazoa”。GeneMan 能结合序列相似性搜索和文本搜索，更有利于发现相关序列，即使在数据库入口存在错误和遗漏。

5.7 引物设计、限制图谱和序列编辑

Lasergene 系统包括其他 3 个应用，简述如下。

PrimerSelect 是设计和分析寡核苷酸的工具，这些寡核苷酸包括 PCR、测序、探针杂交和转录的引物。利用 DNA、RNA 和逆向翻译的蛋白质作为模板，PrimerSelect 可详述退火反应的热动力学特性，鉴定所有可能的引物，并按特定条件下适合的程度排列。对标准 PCR 和多重 PCR 实验，PrimerSelect 也会高亮化潜在的缺陷。

MapDraw 产生限制图谱并以 6 种不同的格式(包括环状和线性图)显示位点、翻译产物、序列特征。输入的序列可以小如寡核苷酸，大至最大的 BAC 插入物。从近 500 个限制位点的数据库中，可选择任何子集来作图。利用布尔运算符组合限制位点，可为辅助克隆策略提供有力的位点选择工具。

所有 Lasergene 系统都提供 EditSeq 以利于操作不同格式的核酸和蛋白质序列，它们包括：GeneMan、GenBank、FASTA、文本格式、ABI、ALF、Staden、剪切板、GCG、MacVector™，以及 1982 年建立的 Lasergene 序列文件格式 DNASTAR。此外，还可利用互联网通过 NCBI 数据库的登录号获得序列，在 EditSeq 中一个序列的相关序列可利用其整合的 BLAST 搜索工具鉴定。EditSeq 提供一些基本的分析工具，包括编辑、反向互补、翻译、逆向翻译、可读框鉴定及简单的注释。

5.8 概要

Lasergene 的 8 个模块使用户能完成序列分析的每个步骤，包括从整理组装序列数据到发现基因、注释、基因产物分析、序列相似性搜寻、序列比对、进化树分析、寡核苷酸引物设计、克隆策略、结果发表等。Lasergene 软件包提供函数和所需的定制工具，使用户完成软件作者未曾想到的分析。

致谢

感谢我的同事 Carolyn Allex 和 Sharon Savage 评阅手稿以及 Jeff Briganti 对 Lasergene 各种性质的耐心讲解。

(吴东林 译)

参 考 文 献

[1] Blattner, F. R., Plunket III, G., Bloch, C. A., Perna, N. T., Burland, V., Riley, M., Collado-Vides, J., Glasner, J. D., Rode, C. K., Mayhew, G. F., Gregor, J., Davis, N. W., Kirkpatrick, H. A., Goeden, M. A., Rose, D. J., Mau, B., and Shao, Y. (1997) The complete genome sequence of *Escherichia coli* K12. *Science* **277,** 1453-1462.

[2] Burland, V., Daniels, D. L., Plunkett III, G., and Blattner, F. R. (1993) Genome sequencing on both strands: the Janus strategy. *Nucleic Acids Res.* **21,** 3385-3390.

[3] Allex, C. F., Baldwin, S. F., Shavlik, J. W., and Blattner, F. R. (1997) Increasing consensus accuracy in DNA fragment assemblies by incorporating fluorescent trace representations, in *Proceedings of the Fifth International Conference on Intelligent Systems for Molecular Biology* (Gaasterland, T., Karp, P., Karplus, K. Ouzounis, C., Sander, C., Valencia, A.). AAAI Press, Menlo Park, pp. 3-14.

[4] Borodovsky, M. and McIninch, J. (1993) GeneMark: parallel gene recognition for both DNA strands. *Comp. Chem.* **17,** 123-133.

[5] Trifonov, E. N. and Sussman, J. L. (1980) The pitch of chromatin DNA is reflected in its nucleotide sequence. *Proc. Natl. Acad. Sci. USA* **77,** 3816-3820.

[6] Chou, P. Y. (1990) Prediction of protein structural classes from amino acid composition, in *Prediction of Protein Structure and the Principles of Protein Conformation* (Fasman, G. D., ed.), Plenum, New York, NY, pp. 549-586.

[7] Garnier, J., Osguthorpe, D. J., and Robson, B. (1978) Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. *J. Mol. Biol.* **120,** 97-120.

[8] Deléage, G. and Roux, B. (1987) An algorithm for protein secondary structure prediction based on class prediction. *Protein Eng.* **1,** 289-294.

[9] Parry, D. A. (1982) Coiled coils in α-helix-containing proteins: analysis of the residue types in the heptad repeat and the use of these data in the prediction of coiled coils in other proteins. *Biosci. Rep.* **2,** 1017-1024.

[10] Engelman, D. M., Steitz, T. A., and Goldman, A. (1986) Identifying nonpolar transbilayer helices in amino acid sequences of membrane proteins. *Annu. Rev. Biophys. Chem* **15,** 321-354.

[11] Kyte, J. and Doolittle, R. F. (1982) A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.* **157,** 105-132.

[12] Hopp, T. P. and Woods, K.R. (1981) Prediction of protein antigenic determinants from amino acid sequences. *Proc. Natl. Acad. Sci. USA* **78,** 3824-3828.

[13] Eisenberg, D., Weiss, R. M., and Terwilliger, T. C. (1984) The hydrophobic moment detects periodicity in protein hydrophobicity. *Proc. Natl. Acad. Sci. USA* **81,**140-144.

[14] Bairoch, A., Bucher, P., and Hofmann, K. (1997) The PROSITE database, its status in 1997. *Nucleic Acids Res.* **25,** 217-221.

[15] Margalit, H., Spouge, J. L., Cornette, J. L., Cease, K. B., Delisi, C., and Berzofsky, J. A. (1987) Prediction of immunodominant helper T cell antigenic sites from the primary sequence. *J. Immunol.* **138,** 2213-2229.

[16] Jameson, B. A. and Wolf, H. (1988) The antigenic index: a novel algorithm for predicting antigenic determinants. *Comp. Appl. Biosci.* (now *Bioinformatics*) **4,** 181-186.

[17] Sette, A., Buus, S., Appella, E., Smith, J. A., Chesnut, R., Miles, C., Colon, S. M., and Grey, H. M. (1989) Prediction of major histocompatibility complex binding regions of protein antigens by sequence pattern analysis. *Proc. Natl. Acad. Sci. USA* **86,** 3296-3300.

[18] Rothbard, J. B. and Taylor, W. R. (1988) A sequence pattern common to T cell epitopes. *EMBO J.* **7,** 93-100.

[19] Emini, E. A., Hughes, J., Perlow, D., and Boger, J. (1985) Induction of hepatitis A virus-neutralizing antibody by a virus-specific synthetic peptide. *J. Virol.* **55,** 836-839.

[20] Martinez, H. M. (1983) An efficient method for finding repeats in molecular sequences. *Nucleic Acids Res.* **11,** 4629-4634.

[21] Wilbur, W. J. and Lipman, D. J. (1983) Rapid similarity searches of nucleic acid and protein data banks. *Proc. Natl. Acad. Sci. USA* **80,** 726-730.

[22] Lipman, D. J. and Pearson, W. R. (1985) Rapid and sensitive protein similarity searches. *Science* **227,** 1435-1441.

[23] Hein, J. (1990) Unified approach to alignment and phylogenies. *Meth. Enzymol.* **183,** 626-645.

[24] Higgins, D. G. and Sharp, P. M. (1989) Fast and sensitive multiple sequence alignments on a microcomputer. *Comp. Appl. Biosci.* (now *Bioinformatics*) **5,** 151-153.

[25] Pearson, W. R. (1990) Rapid and sensitive sequence comparison with FASTP and FASTA. *Meth. Enzymol.* **183,** 63-98.

[26] Bucher, P. and Bairoch, A. (1994) A generalized profile syntax for biomolecular sequence motifs and its function in automatic sequence interpretation, in *Proceedings 2nd International Conference on Intelligent Systems for Molecular Biology* (Altman R., Brutlag D., Karp P., Lathrop R., Searls D., eds.), AAAI Press, Menlo Park, CA, pp. 53-61.

6 PepTool™和 GeneTool™：非平台依赖性的生物序列分析工具

David S. Wishart Paul Stothard Gary H. Van Domselaar

6.1 引言

PepTool™和 GeneTool™是两个由 BioTools 公司(www.biotools.com)提供的新的生物信息学软件包。从它们的名称上看，PepTool 用于蛋白质序列分析，而 GeneTool 用于 DNA 序列分析。合并的软件包对学术用户价格为 1500 美元，对商业用户价格为 1875 美元。PepTool 实际上基于两个公共程序 SEQSEE[1]和 XALIGN[2]，它们最初由 Alberta 大学开发。这两个基于 UNIX 系统的程序后来被移植到其他平台上，生成图形用户界面[3]，后来授权给 BioTools 作为商用软件包，叫做 PepTool。GeneTool 由 BioTools 独立开发，但它采用了 PepTool 的一些关键概念和算法。PepTool 1.0 版发行于 1997 年 12 月，GeneTool 1.0 版发行于 1998 年 12 月。

PepTool 和 GeneTool 都是综合和集成的程序，能提供全方位的分析和图形性质，这些性质常见于一些高级的生物信息学产品。PepTool 和 GeneTool 给生物信息学的许多方面带来了必要的进步，如算法设计、图形界面应用、数据压缩、网络并行技术、互联网通讯等。但其最重要的革新在于 PepTool 和 GeneTool 都是非平台依赖性的软件包。这意味着这两个程序能运行于任何计算机或任何操作系统上(MacOS、Windows、UNIX)，而程序的整体外观和感觉没有明显变化。BioTools 之所以能做到这点，是因为 PepTool 和 GeneTool 的图形用户界面(GUI)是用一种特殊的语言 Smalltalk 开发的。Smalltalk 是 20 世纪 70 年代早期由 Xerox 的 Palo Alto 研究中心开发的，本质上是众所周知的平台非依赖性 GUI 程序设计语言 Java 的复杂版本。Smalltalk 允许制作复杂的 GUI 而不必考虑平台兼容性和底层(back-end)设计，也不必考虑较慢的程序编译器对工作的限制。

以下我们将试图突出介绍 PepTool 和 GeneTool 的一些有用的特征。重点放在每个程序特殊的组件上。由于篇幅限制，我们不能对两个软件包进行完整的概述，但我们希望这个简短的介绍能使读者对这些强大的生物序列分析工具的设计、目的、基本用途有一些了解。

6.2 方法和系统需求

所有 PepTool 和 GeneTool 的复杂分析功能(如 back-end)是用 ANSI C 写的。GUI 和一些简单分析功能都是用 VisualWorks Smalltalk(2.5.2 版，ObjectShare)写的。PepTool 和 GeneTool 支持 Power Macintosh(OS7.5 版或更高)，Windows 兼容 PC(Win95、Win98 和 WinNT)，Silicon Graphics(Irix 5.0 版或更高)和 Sun(Solaris 2.0 版或更高)等平台。用于其他操作系统的版本需要特别定购。网络并行(见下文)适用于 SUN、SGI 和 Windows 机器，也适用于 1999 年晚期的 Mac 机。无数据库的联合软件包(PepTool 和 GeneTool)需要 70M 的磁盘空间(25M 的 PepTool 和 45M 的 GeneTool)。每个软件包都同时发行自己的序列数据库，而用户也可购买不同的升级选择。特殊压缩的 PepTool 蛋白质数据库需要 60M，而压缩的 GenBank 数据库需要 3.5G。运行 PepTool 和 GeneTool(加数据库)的计算机至少需要 32M RAM(建议 64M)和 4G 的自由磁盘空间。PepTool 和 GeneTool 在兼容 PostScript 的打印机和运行于 Windows 下的打印机上都支持所见即所得的打印。

6.3 PepTool 特有的程序特征

因为通用平台兼容性的要求，PepTool 的 GUI 并不严格依附于任何单一的 OS(操作系统)界面协议，但通常它在很多方面更趋近 MacOS 的风格。根据所处平台的不同，程序可能起始于 Finder 或 Multifinder(MacOS)，通过点击 Windows 开始按钮(Win95/98/NT)，或通过输入 PepTool(UNIX)。起始后，应用程序“启动平台”(“launcher”)(图 6.1)出现在屏幕上方，同时一个序列编辑器窗口出现在屏幕中央。PepTool 的启动平台允许用户启动其他窗口，进入

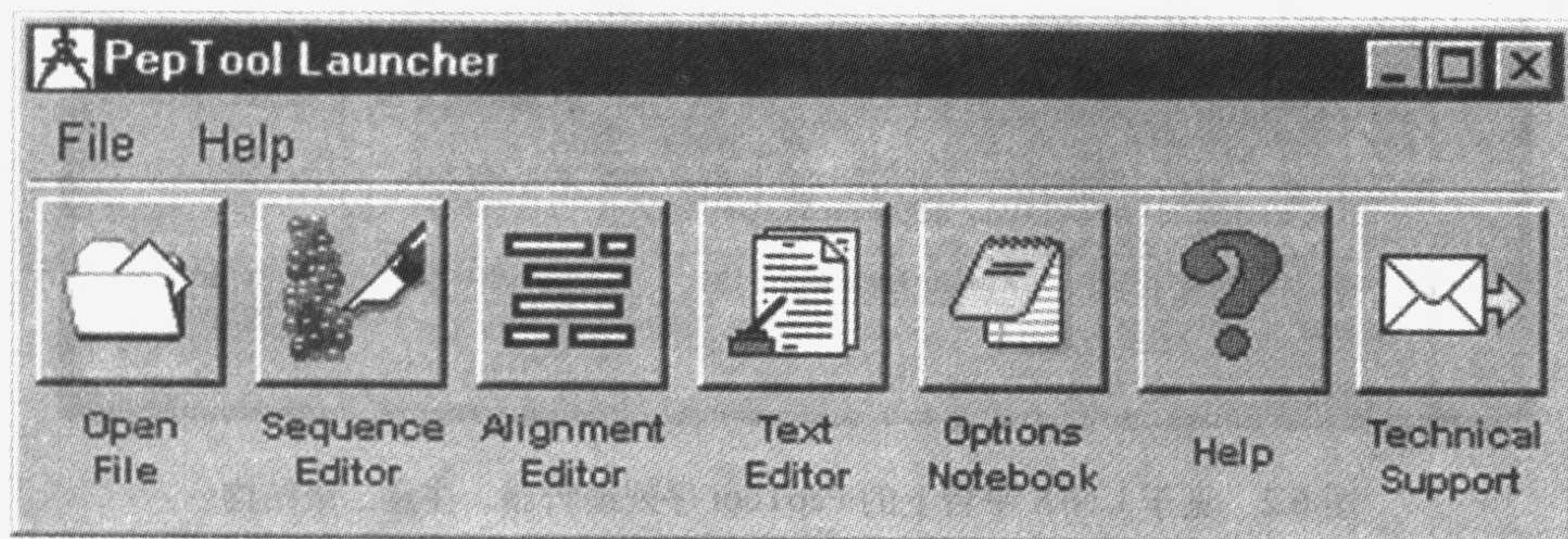


图 6.1　见于 Win95 平台上的 PepTool 应用程序起始平台。从这个程序起始平台可进入许多不同的应用程序和窗口

帮助文件，改变程序参数或通过电子方式联系 BioTools 公司。实际上，PepTool 可通过 PepTool 启动平台或序列编辑器进入许多不同的视窗或窗口，包括：序列编辑器、比对编辑器、简单文本编辑器、图形视窗/编辑器、DotPlot 视窗/编辑器、Helical Wheel 视窗/编辑器、结构视窗/编辑器、序列基序视窗/编辑器、序列统计视窗、帮助视窗、参数编辑器、错误(bug)报告器。这些不同窗口产生的文本文件、文件夹或图形文件可以以窗口专一的格式保存或自动标记(图标和 3 字节扩展名)。所有 PepTool 文件、文件夹和目录都可用文件选择器(与用户系统专一性的文件搜索器类似)搜索或浏览。

6.3.1 序列编辑器

序列编辑器(图 6.2)的功能是作为输入、标记、检索、图形或分析蛋白质序列的一个中心工作界面(workspace)。所以大多数 PepTool 的功能是通过这个特殊窗口进入的。序列编辑器包含一系列标准菜单项，包括：File(文件处理和打印功能)、Edit(编辑所见序列)、Transfer(转移序列或所选部分到其他应用程序或窗口)、Search(从数据库中发现或检索序列)、Analyze(完成统计或结构预测)、Graph(描绘物理化学性质或序列相似性)和 Help(进入依赖上下文关联的超连接帮助系统)。

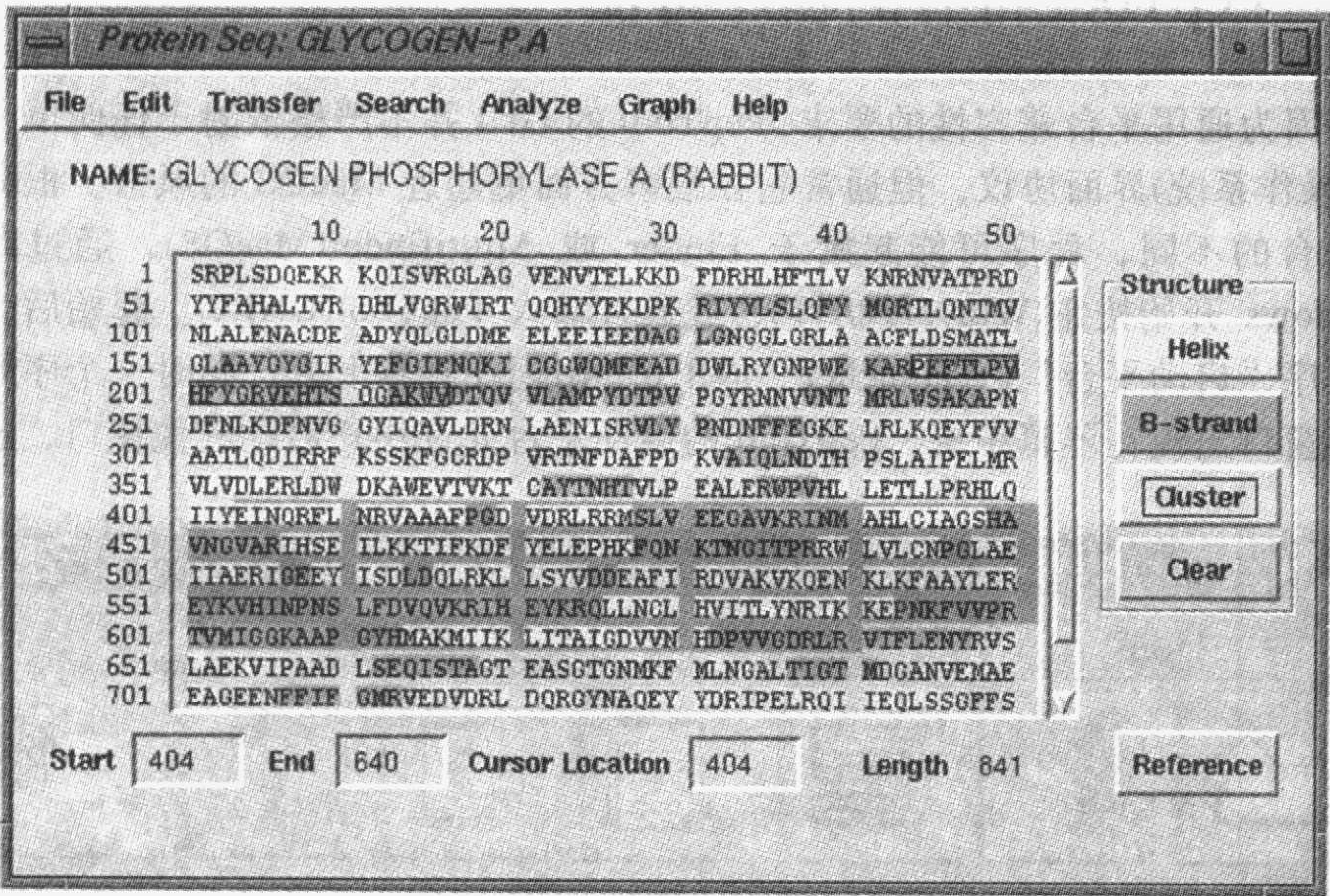


图 6.2　见于 UNIX 平台上的 PepTool 序列编辑器。注意二级结构是如何被标记和观察的

自动装载或手动输入序列编辑器的序列可保存为 SwissProt、PIR、PepTool 或 ASCII 格式。编辑器也能阅读外来格式的文件，如 GCG、IntelliGenetics、FASTA、

SwissProt 和 NBRF-PIR，以及其他通用文件形式。外来格式阅读器是智能化和通用的，即它不需要用户知道或预设一个给定的序列文件格式。同样，如果外来格式阅读器遇到一个其从未见过的文件格式，它通常能做出适当的选择，从多余的文本中分解出序列。

与大多数序列编辑器一样，PepTool 的序列编辑器支持自动空格、自动换行和鼠标驱动的文本选择，以进行通常的剪切、粘贴、拷贝和段落删除操作。它还有文本输入过滤器(当键盘输入非 IUPAC 字符时屏幕会闪烁)、序列标尺、实时序列长度监视器和可编辑的光标位置框，后者当鼠标单击或文本输入操作导致光标位置变化时可立即更新。序列和序列文件的信息显示在窗口顶端，而其他数据(如登录号、杂志索引、日期等)可从一个弹出的序列参考卡片中读取和输入(通过窗口右下角的索引按钮进入)。

PepTool 序列编辑器的一个尤为有用的性质是支持彩色标注的二级结构显示和编辑。位于窗口右侧的一些按钮使用户在完成配对序列的比对时能直接在序列上画出二级结构(如果已知)或预先把一些残基集合在一起。当观看从 PepTool 结构数据库(包括数百个已知二级结构的序列)装载的序列时，这些按钮还能作为彩色标注的图例。

6.3.1.1 数据库搜寻

PepTool 允许从不同数据库进行多种形式的序列数据库搜寻，所有搜寻都起始于序列编辑器(位于 Search 菜单项)。利用数据浏览器可查看、保存、传递数据库搜寻的结果(图 6.3)。PepTool 支持数据库查询和序列检索(retrieval)，这种查询可基于关键词(如生物体、蛋白质名称、登录号、partial name 或它们的逻辑组合)，序列模式(pattern)(简单序列片段和复杂序列模式)，序列相似性[短的相似序列串(stretches)]，尤其是全局(global)序列同源性。PepTool 提供两种全局序列同源性搜寻的选择——快速搜寻和详细搜寻。

FASTALIGN[1](快速搜寻)，在个人计算机上所耗的时间一般少于 5min，它基于与 FASTDB、FASTA 和 BLAST 类似的技术，尽管它采用了专门开发的评分矩阵(scoring matrix)，并生成了全局比对(global alignment)以取代部分本地比对(partial local alignment)(BLAST 通常这样做)。FASTALIGN 对 FASTDB 的并列比较表明 FASTALIGN 比 FASTDB 稍快而且更灵敏[1]。

NWALIGN[1](详细查询)，在个人计算机上(无网络并行)一般耗时数小时，它基于 Needleman-Wunsch 算法[4]。独立检验显示这是一个强大的远程序列鉴定算法，它的表现远远优于 BLASTP、BLITZ、DFLASH 或 FASTA[5]。有趣的是，用于 PepTool 的相同算法在鉴定新型痘病毒编码的病毒蛋白质[6]和新的尿嘧啶糖基化酶(glycosylase)[7]时发挥了重要作用。

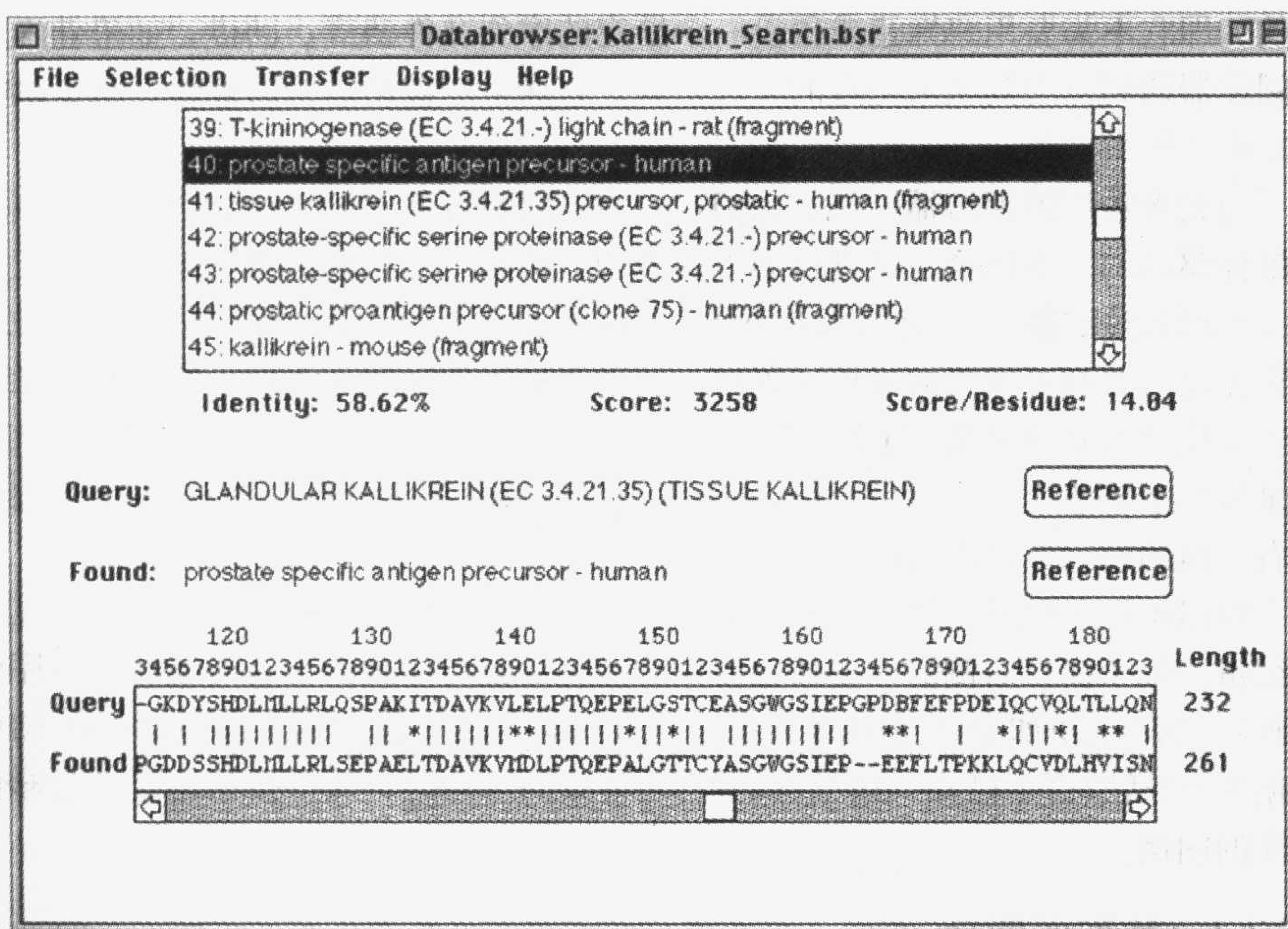


图 6.3　见于 Power MacOS 平台的 PepTool 数据浏览器。上方窗口可选择命中的数据，底部窗口可查看比对

6.3.2　比对编辑器

比对编辑器(图 6.4)是一个直观的工具，它可以查看、编辑并自动生成配对和多重序列比对。一般数据是从数据浏览器和序列编辑器传递到本窗口的。从编辑菜单，用户能很容易添加或删除特定序列，或者改变给定的序列或序列名。当序列被载入和编辑后，通过点击右下方的 Compute Alignment 按钮可自动计算序列比对。PepTool 采用 XALIGN 算法[2]完成这个操作，在比对过程中，能利用序列聚类(clustering)和二级结构信息迅速排列数百个序列。在比对视窗上方的窗口中可按照一定的阈值生成共有序列，该阈值显示在共有序列阈值框(consensus threshold box)中。在 Display(显示)菜单下，用户可选择比对的色彩显示方式，选项包括按照结构(双色)、特性(property)(12 色)或一致性(identity)(单色)。从显示菜单项也可计算和观察配对的比较矩阵(matrix)。

也可通过在序列块(block)上选择或上色(painting)，从而对自动生成的比对进行手工比对和编辑。当包括一个或多个部分序列的全部序列块被加亮后，可用窗口下方鼠标激活的箭头进行左右移动。

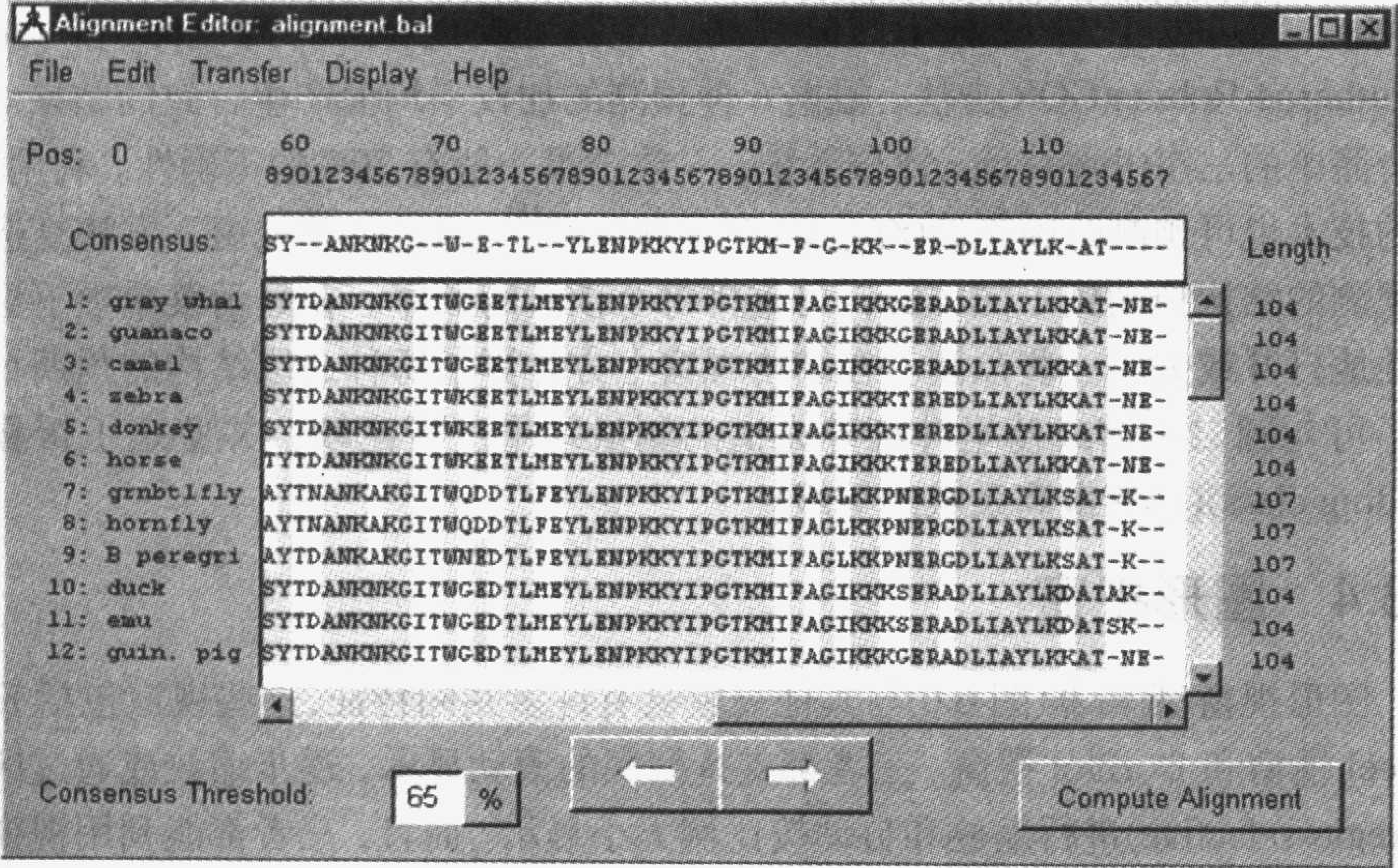


图 6.4　见于 Win95 平台上的 PepTool 比对编辑器。根据序列同源性可查看或彩色标注多序列比对

6.3.3　结构视窗

结构视窗(图 6.5)显示了预测的二级结构，用特殊阴影和色彩代表螺旋和

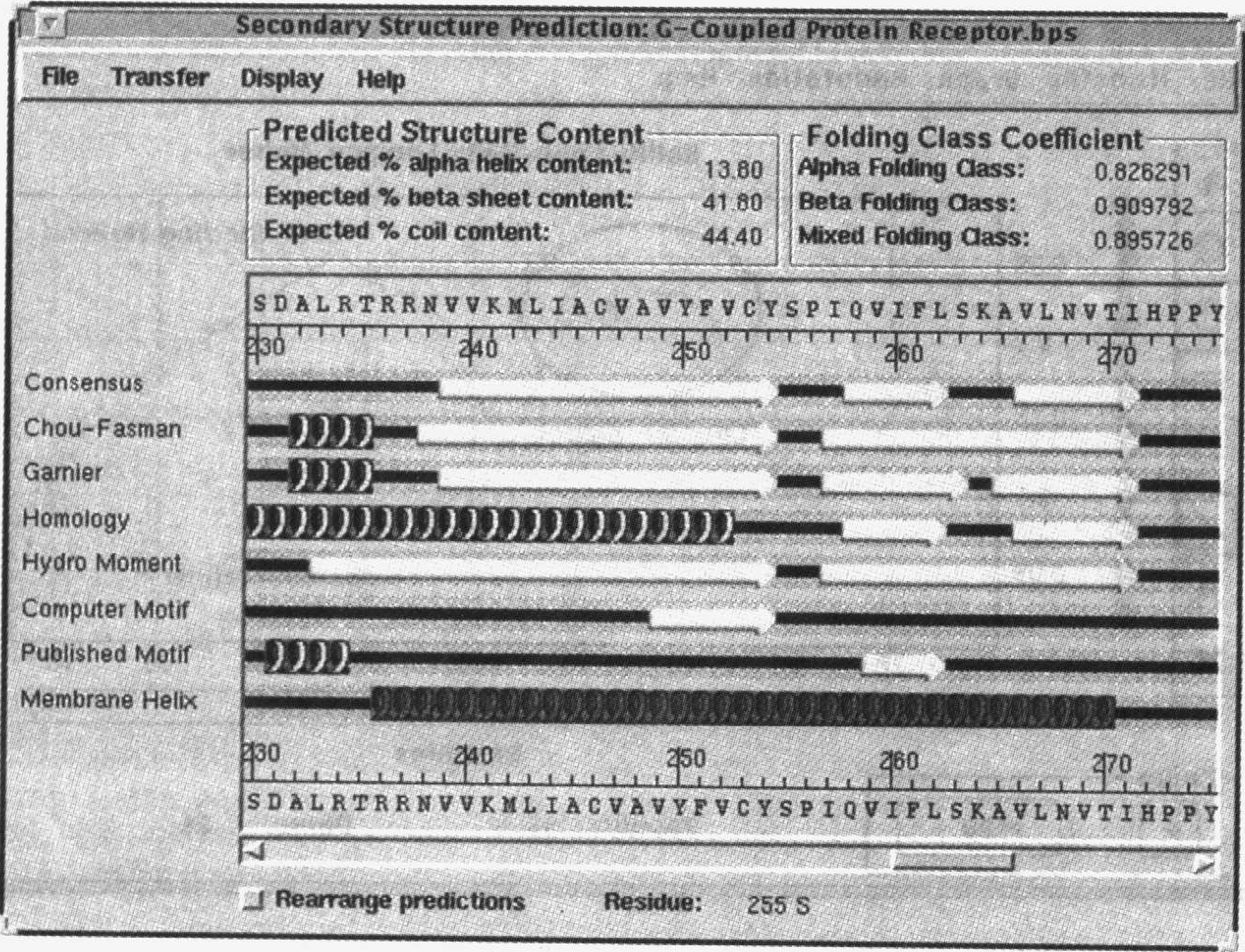


图 6.5　见于 UNIX 平台上的 PepTool 结构视窗。线圈表示螺旋而箭头表示β折叠

β折叠。共可以生成 6 种不同的预测，包括经典的 Chou-Fasman 和 Garnier-Osguthorpe-Robson(GOR)方法。根据 6 种预测的加权平均值产生一致的结果。基于一个简单的三态(three-state)评分系统，一致结果一般有 70%的正确性。利用 Klein 等的技术也可预测跨膜螺旋(红色)的存在和位置[8]。通过打开视窗下方的复选框并把预测的结构拖动到不同位置，可重排每个预测的顺序。在显示菜单项下也可选择性地打开或关闭一些预测。在结构视窗上方计算了每种二级结构的预期百分含量，可与圆二色谱或傅里叶红外光谱的测量结果进行比较，并鉴定了折叠等级(folding class)(具有最高系数的等级)。

6.3.4 图形视窗

图形视窗/编辑器(图 6.6)与其他视窗具有许多共同特征，这些视窗包括 Helical Wheel 视窗和 DotPlot 视窗。三者都支持彻底的滚动显示、逐步缩放或通过区域选择(regio-selective)进行缩放和自动按比例大小显示。而且，三者都能利用视窗左侧的图形面板在显示图形上添加或删除文本、线条、箭头、框或圆。精心设计的图形视窗可显示疏水性、疏水力矩(hydrophobic moment)并预测柔性。利用图形菜单可进一步编辑蛋白质性质图形，用户可调整图形颜色、线条宽度、图形标题和坐标轴名称等，还可打开或关闭网格线和残基标志。通过注释菜单，也可交互式地选择和调整任何图形注释(除文本外)的颜色、线宽和线型。

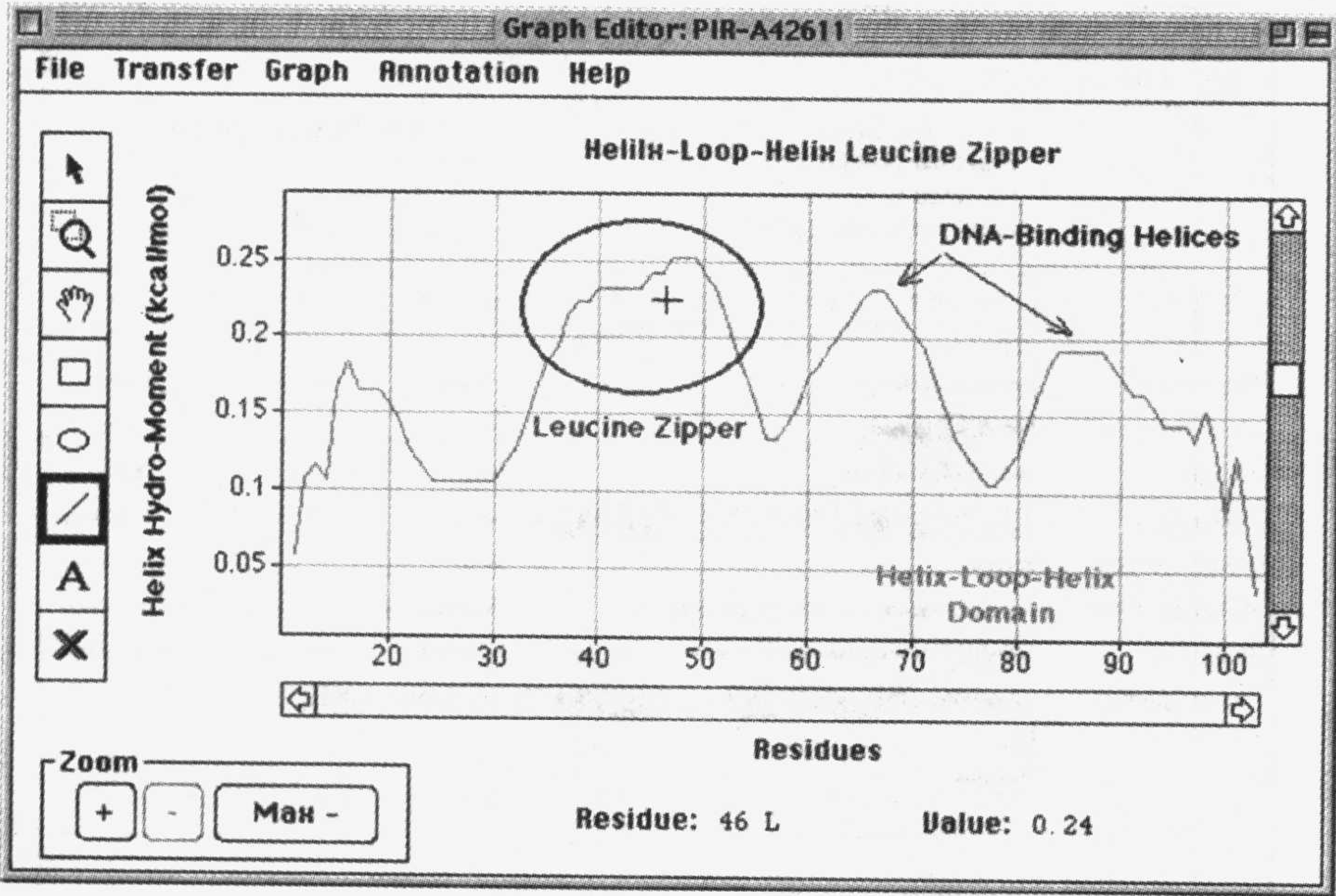


图 6.6 见于 Power MacOS 平台的 PepTool 图形视窗/编辑器。阐明了对螺旋的疏水力矩的注释曲线

6.3.5 DotPlot 视窗

利用 PepTool 的 DotPlot 视窗(图 6.7)可显示、编辑、注释和评估 Dot 矩阵或 DotPlot 序列对比。可进行 2 个不同序列的配对比较和简单的自身序列比较。通过可编辑的“严谨性”、“窗口大小”和“斜线过滤器”框可调整所绘斜线的数量和长度。同样通过图形菜单内的选项也可改变图形的颜色，以及坐标轴和图形的标题。在显示序列相似性水平时，PepTool 的 DotPlot 程序采用了一种独特的、简化的色彩明暗显示方案显示序列相似性的水平，完全匹配的最亮，而匹配差的逐渐变浅。DotPlot 视窗与 PepTool 中其他图形视窗一样，允许缩放和注释操作，但与其他不同的是，它允许在底部的序列窗口中查看选定斜线所代表的序列。其方法是，首先点击注释面板上的 ATGC 按钮，然后点击 DotPlot 窗口中的特定斜线。对应于该斜线的配对序列比对将以高亮度方式出现在底部窗口中，以便查看。

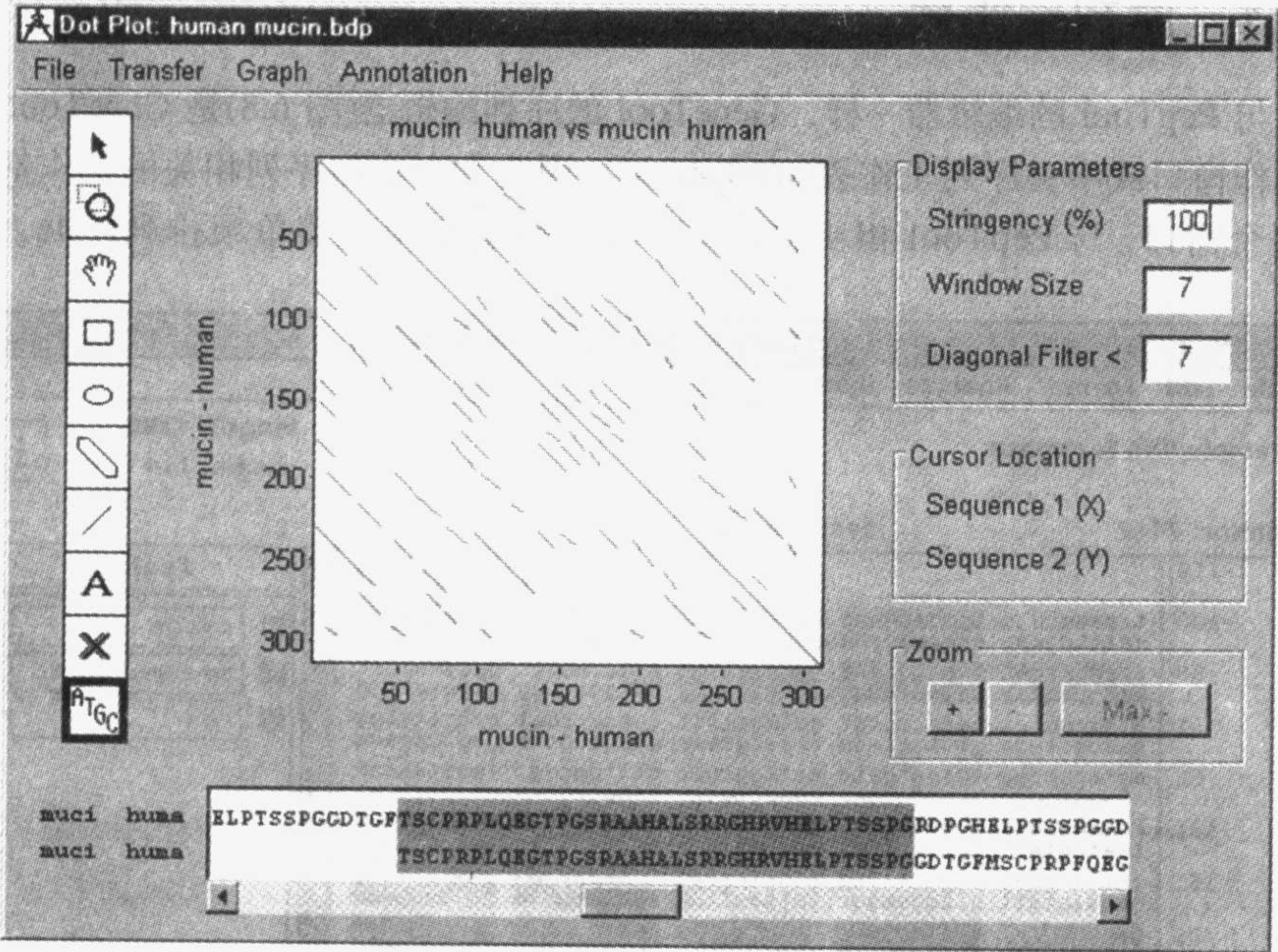


图 6.7　见于 Win95 平台上的 PepTool 的 DotPlot 视窗。强的斜线表示在该蛋白质中存在多个内部重复

6.4 GeneTool 的程序特性

GeneTool 与 PepTool 共享许多设计和版面(layout)特征，但它也有大量重大改进(其中许多改进可能会用于 PepTool 的 2.0 版)。特别是，GeneTool 支持可调节大

小的窗口、可调节大小的字体、多特征显示、多特征编辑、打印预览的注释和音频回放。它也可处理数据搜索、参数选择、参考信息、窗口缩放，并以更直观的方式管理窗口。与 PepTool 一样，GeneTool 可起始于 Finder 或 Multifinder(MacOS)，点击 Windows 的起始按钮(Win95/98/NT)或输入 GeneTool(UNIX)。起始后 GeneTool 的平台(launcher)出现在屏幕上方，同时序列编辑器窗口出现在屏幕中央。通过 GeneTool 的起始平台和序列编辑器，可进入超过 20 个不同的视窗和窗口，包括：GeneTool 序列编辑器、翻译视窗/编辑器、色谱(chromatogram)视窗/编辑器、比对(alignment)编辑器、重叠群(contig)编辑器、版面(layout)或介绍(presentation)编辑器、图形视窗/编辑器、限制图谱视窗/编辑器、特征(feature)/外显子/序列基序视窗/编辑器、PCR 引物设计程序、凝胶模拟视窗、序列统计视窗、帮助视窗、参数选择编辑器和错误报告程序。利用与 PepTool 类似的文件选择器(file chooser)可搜寻或浏览所有 GeneTool 的文件、文件夹、目录。

6.4.1 序列编辑器

与 PepTool 的编辑器一样，GeneTool 的序列编辑器(图 6.8)是 GeneTool 的中央操作窗口或中央序列工作表(worksheet)。因此大多数与序列相关的操作都开始于这个窗口。与 PepTool 相比，GeneTool 保持了相同的菜单项排列(File、Edit、

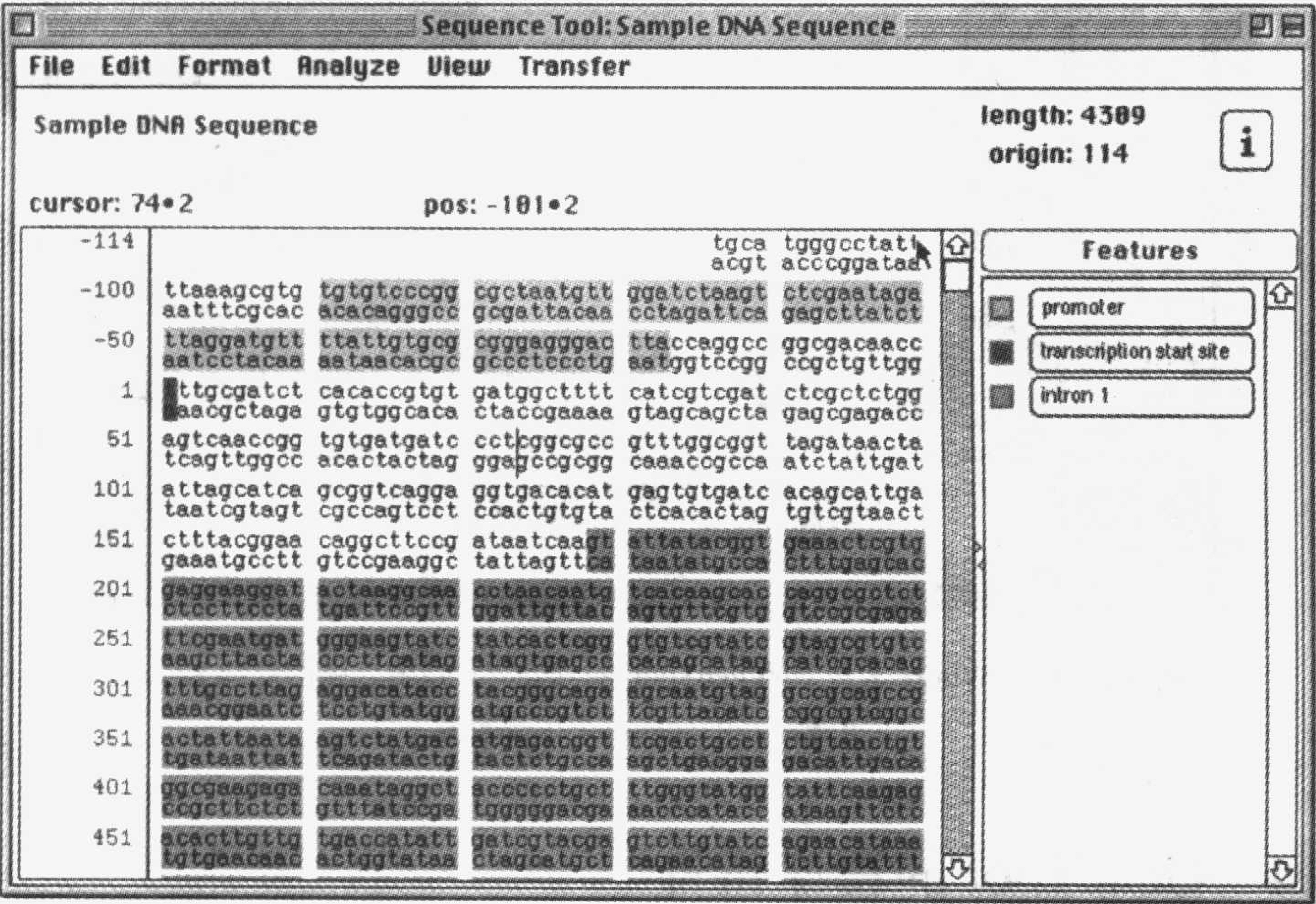


图 6.8 见于 Power MacOS 平台的 GeneTool 序列编辑器。
注意序列特征被标注和查看的方式

Format、Analyze、View、Transfer)，允许选择阅读或保存相同的序列格式，包括 EMBL、GenBank、日本 DNA Date Bank(DDBJ)格式。为了限制 PepTool 文件类型的增加，GeneTool 的设计者将从一种给定的序列分析得到的多种文件类型统一成单一的序列文件。与给定的序列文件相关的图形、图表、模拟或其他分析功能都可利用 View(视窗)菜单选择和观察。与大多数其他 DNA 序列编辑器一样，GeneTool 编辑器允许不同的字符集合(character grouping)(1，3，5，10 等)，单链或双链显示，DNA 转换成 RNA、链反向、链互补、大小写显示、音频回放、自动调整间距、自动回折，以及利用鼠标驱动的文本选择进行剪切、粘贴、拷贝和段落删除操作。它还支持简并的 DNA 字母表(当从键盘输入非 IUPAC 字符时会闪烁)，还支持不断升级的序列长度、可读框和光标位置框。

与 PepTool 的编辑器相比，GeneTool 的序列编辑器在版面和设计上有很多不同。尤其是序列名称、序列长度、光标位置、读框出现在序列状态条上。另外，参考信息(reference information)按钮被移到上方并用一个“I”图标代替，它允许显示更广泛的注释和参考。GeneTool 的序列编辑器最明显的变化可能是它利用一个可编辑的滚动“特征图例”框来支持复杂的特征显示和标记(markup)系统。利用这个系统，可装载 GenBank、EMBL、DDBJ 序列，而且可通过色彩标记的文本选择器自动显示它们的特征表。在这个可扩展的特征图例框中可观察到特征的全名以及其相应的颜色。利用与每个 Feature Name(特征名称)按钮相连的彩色单选按钮可关闭和打开在文本窗口中以颜色表示的每个特征。通过按下 shift 键同时点击特征名称按钮(或点击图例框上方的特征按钮)，可显示包含有关此特征(或所有其他特征)的其他信息的对话框。这个对话框使用户能添加、记录、编辑、注释或按次序排列(prioritize)重叠的特征。这个特征表示(rendering)方法的一个关键优势是允许用户向新的序列数据添加他们自己的特征(PepTool 允许在其编辑器中以类似的方式添加或去除二级结构)。这可通过以下步骤来简单完成：向图例上加入新的特征按钮或把一个已有特征按钮编辑成预想的特征名称；在文本窗口中高亮化特征序列；并点击相应的特征按钮给高亮化的文本标记颜色。

6.4.2 色谱视窗(chromatogram viewer)

利用 GeneTool 的色谱视窗可以不同的格式对从 DNA 自动测序仪获得的未加工的(raw)序列数据进行读取、编辑和保存操作(图 6.9)。尤其是，数据可从 ABI 或 SCF 格式的色谱文件及 GeneTool 自己的色谱格式中直接读取。每个色谱曲线(trace)可通过窗口左侧的彩色 A、C、G 和 T 按钮或通过核对 Format(格式)菜单下的 Base Trace(碱基曲线)选项关闭或打开。每个曲线可被选择或上下拖动以利于看清楚碱基。它还具有 5′/3′整理(trimming)特性(位于编辑菜单下)，能消除不需要的或色谱最末端不能读取的数据。所有 4 个色谱曲线的垂直高度比例可用屏幕右侧的比例调节杆来调整。利用测序仪得到的碱基可以用与大多数文本编辑器类似的

方式改变或删除。但是，插入一个或几个碱基必须经过编辑菜单的形态改变。色谱视窗也支持两种形式的 Find(查找)功能，一个被设计用来定位不确定的碱基(Find Next Problem)，另一个用来定位特殊的亚序列(Find…)。

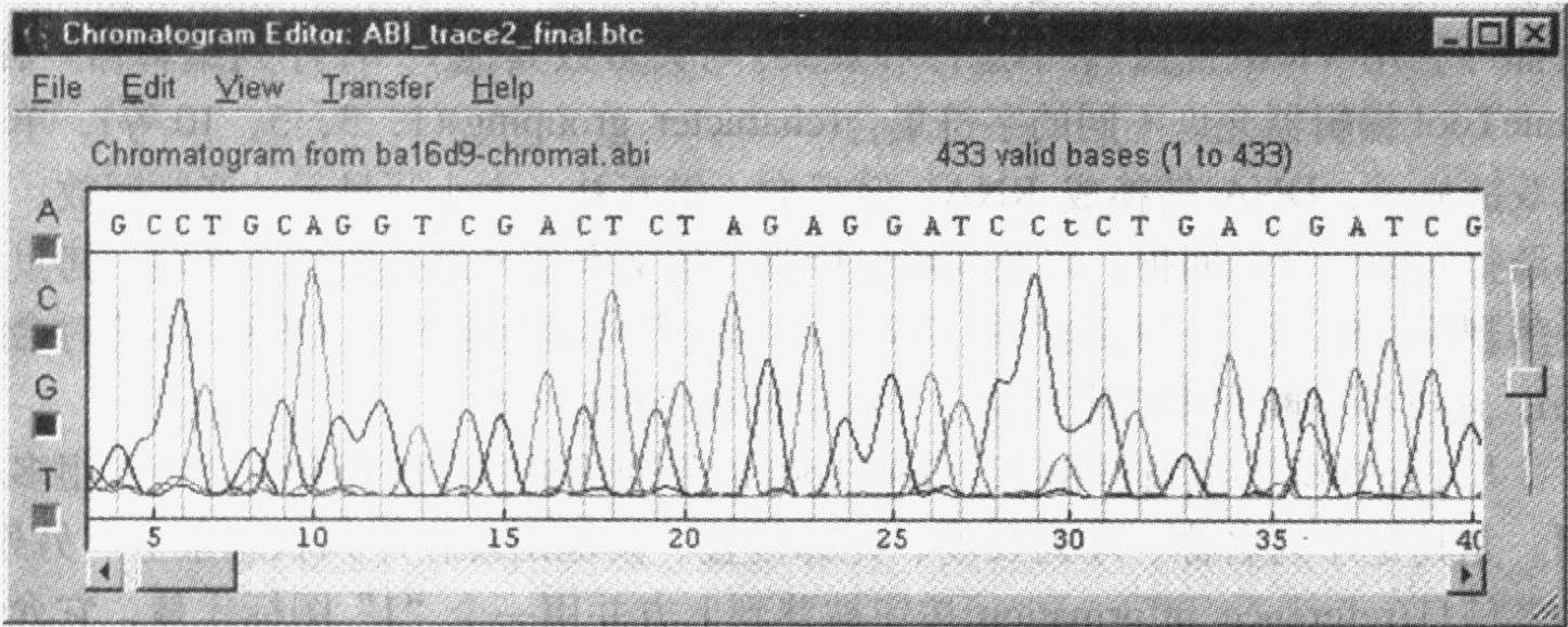


图 6.9　见于 Win95 平台上的 GeneTool 色谱视窗

6.4.3　外显子发现器

GeneTool 利用一个特殊的方法鉴定外显子/内含子在真核生物 DNA 中的位置，它基于参考分数逻辑方法(reference point logistic，RPL)，这个方法是由 Alberta 大学的 Peter Hooper 开发的。RPL 类似复杂的神经网络系统，它可以被训练用来识别复杂的图案和信号，如在外显子/内含子边界发现的信号。利用 Burst 和 Guigo[9] 的检测数据(包含 570 个脊椎动物基因)进行效能评估，表明 RPL 可有效预测外显子和内含子的位置，相关系数超过 0.92(P. Hooper，个人通讯)。它实际上优于大多数其他基因发现算法，包括流行的程序，例如 GRAIL 和 GRAIL2[9]。另外，在标准的桌面机器上 RPL 预测只耗时数秒钟。BioTools 通过添加一种数据库搜寻方法改进了最初的外显子/内含子预测，从而增强了 RPL 技术。尽管这个技术使分析时间增加了 4min，但它可使预测效果提高 3%~4%。

当从 Analyze(分析)菜单中选择 Find Exons/Introns(查找外显子/内含子)选项时，会出现一个对话框，提示用户选择快速搜寻(纯 RPL 法)或详细搜寻(结合 RPL 和快速数据库扫描)。图形化的结果显示在如图 6.10 所示的窗口中。通过鼠标点击可选择单独的外显子或复杂的外显子串，并把它们传送到序列编辑器中，或者利用编辑菜单中的剪接操作使显示的外显子串剪接在一起。值得注意的是这个视窗(以及其他显示基序或特征的图形视窗)允许以各种方式缩放，直到序列水平，所以整个基因序列都能被观察研究。

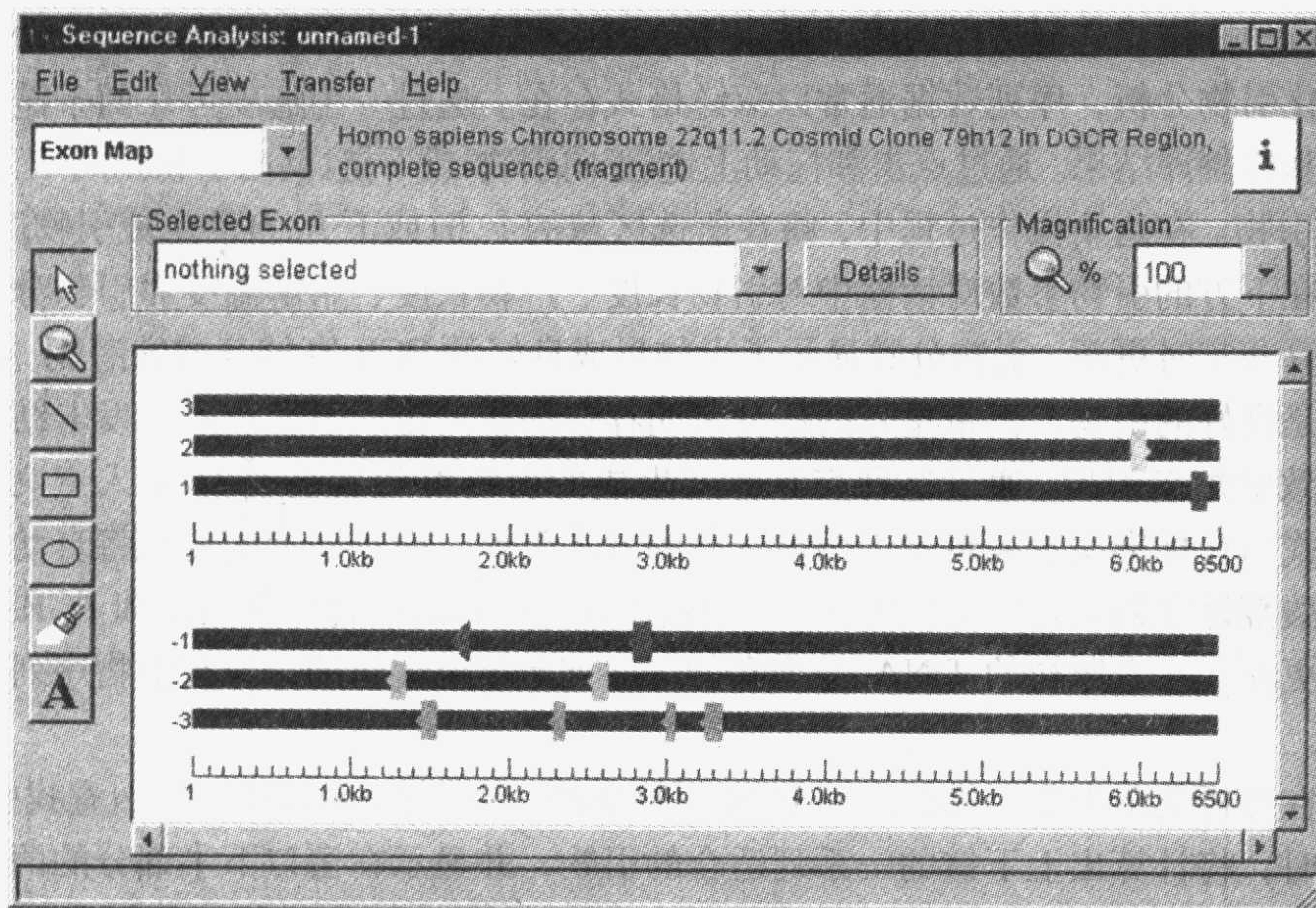


图 6.10　由 GeneTool 的外显子发现器产生的外显子/内含子图谱

6.4.4　PCR 引物设计器

引物设计器(图 6.11)是 PCR 引物选择和设计的工具，它既可以交互方式工作，

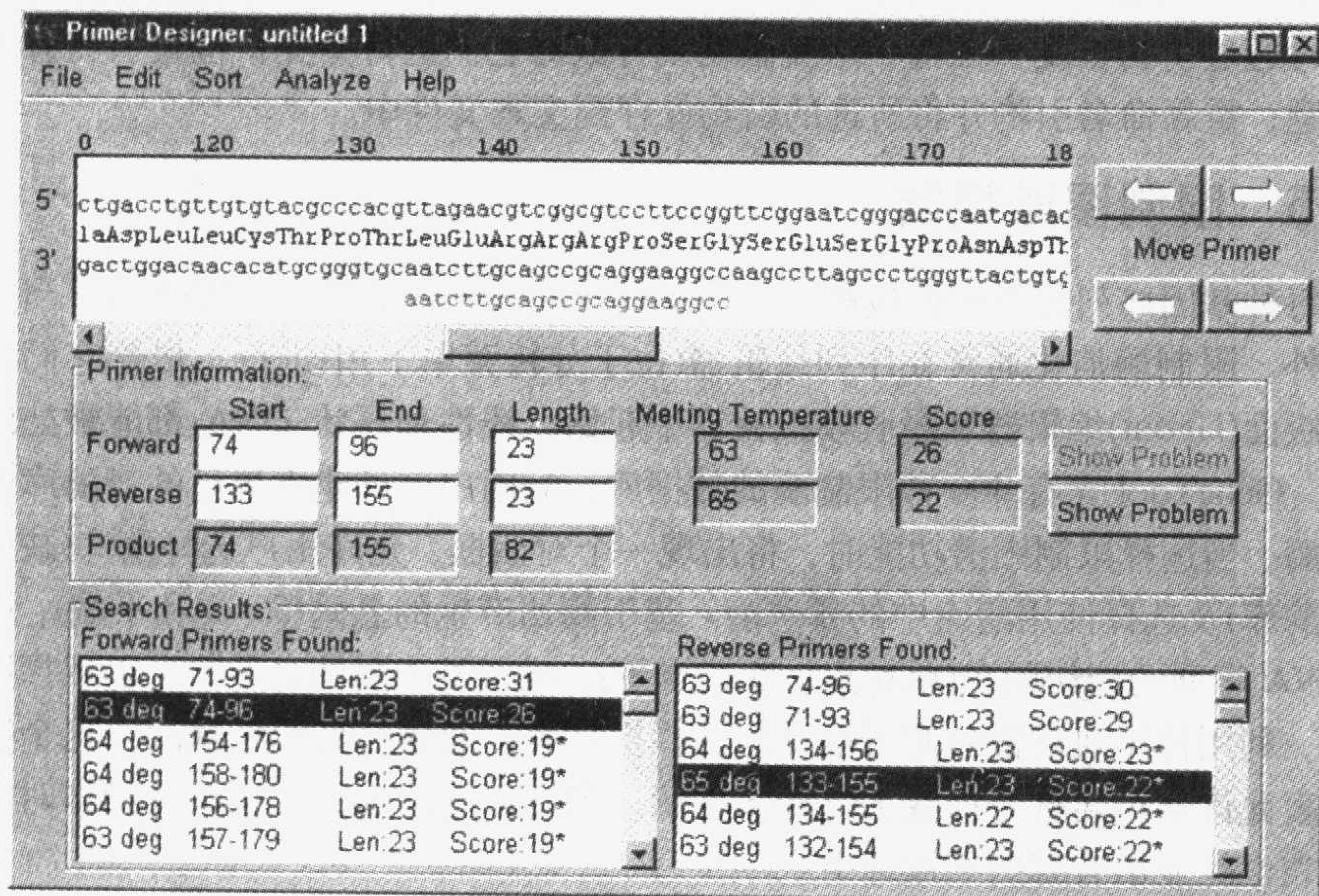


图 6.11　见于 Win95 平台上的 PCR 引物设计器。注意窗口下半部分的可选择引物列表

也可自动操作。它能从序列编辑器或 GeneTool 的起始平台(launcher)中启动。为了简化引物分析，序列数据通常以双链格式存在。通过一个选项可在两条链间显示氨基酸翻译结果。通过点击和拖动上链(正向引物)或下链(反向引物)可手工形成 PCR 引物。在这个操作过程中，将在所选区域的上方(或下方)自动生成引物序列，同时在下方的引物参数对话框中对引物长度、产物长度、熔解温度和引物分数进行计算和实时更新。引物分数显示了引物形成良好的 PCR 寡核苷酸的潜能。高分表明是良好的引物，而带星号的低分表明存在潜在的错误引发位点、发卡结构或不一致的熔解温度。通过这种交互方式生成的引物可随后被编辑(引入点突变)，其方式与在标准文本编辑器中编辑字符的方式类似。引物序列的变化将自动导致氨基酸翻译序列的相应变化(包括颜色变化)并对计算的引物熔解温度和 PCR 分数进行更新。注意原初的 DNA 序列和翻译序列是不可编辑的，只有引物序列可被编辑。

在分析菜单中也可进行自动 PCR 引物选择。当选择 Find Primers…(查找引物)操作时，将计算出上下链的一系列适合的引物，并显示在窗口下半部分的两个数据框中。正向(上游)引物显示在左侧，反向(下游)引物显示在右侧。列表可以滚动，通过鼠标点击可选择单独的引物。以这种方式选择引物可观察引物序列并对引物设计器中央的参数框进行更新。利用上述的相同引物编辑技术，可对这些引物进行编辑、延长或缩短。注意手工或自动方式的 PCR 引物参数都可置于 Primer Parameters 对话框中。

GeneTool 的 PCR 引物设计器也支持以下功能：在上下链中查找序列或亚序列；可通过引物长度、位置、熔解温度或分数对它们进行分类；检查引物有无特殊问题；重新命名引物并将所选的引物保存到文本文件中。

6.4.5 限制图谱视窗

每种基因序列分析软件包都有一些图形化的限制图谱视窗，GeneTool 当然也不例外。限制酶消化通常在序列编辑器(位于分析菜单下)中进行，尽管它们也可以从版面(layout)编辑器和凝胶模拟视窗中起始。线性和环状 DNA 都能被处理和显示。GeneTool 含有 400 种限制酶的数据库，但用户也可建立自己的亚酶库和添加新酶。当选择限制图谱功能时，将出现一个对话框，允许用户选择酶库(缺省值为全酶库)及选择使用酶库中的哪些酶。基于酶切产生的悬垂(5′、3′和平端)，所加工 DNA 序列的酶切频率(单切点、双切点等)，或者基于酶的名称，可以选择特殊的酶。利用位于对话框右侧可滚动的选择框列表，能通过名称选择酶。这个特殊的列表允许通过名称选择或去选择多种酶。当用户进行酶的类型选择操作时，这个列表也能显示有哪些酶被选中。

当完成限制消化后，将生成一个如图 6.12 所示的图形化的图谱。如果先前已鉴定了序列特征，它们将显示为彩色的条带或半圆。点击任意彩色特征，会在窗

口顶端的状态条中显示特征的信息。当激活后，相同的特征将转移到序列编辑器中作进一步分析。除可显示序列特征外，也可显示酶切位点。单限制切点显示为蓝色，而多限制切点显示为黑色。点击任意限制酶标记将出现一个弹出框，显示放大的序列区域，其酶识别序列以高亮度的红色表示。酶标记(以及相关的位点线)可移动或拖动到屏幕的任何位置，使得更易于阅读或以对称的方式显示。点击两个酶的名称，同时按下 shift 键，用户能够选择两个切点间的 DNA 序列。这个图形化的消化片段可以被剪切、拷贝或粘贴到另一个序列或另一个序列编辑器中。

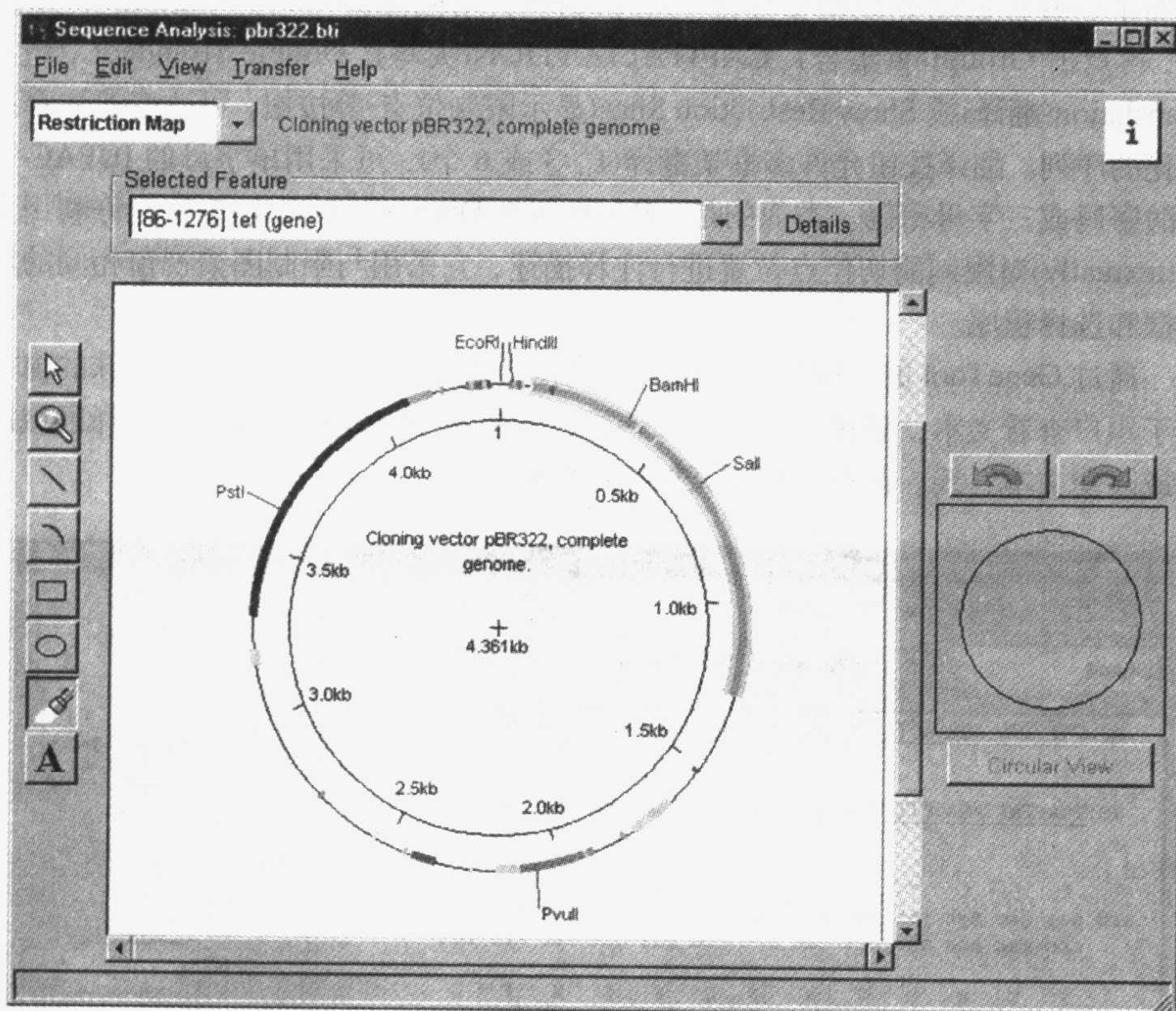


图 6.12　利用 GeneTool 的限制图谱器制作的 pBR322 质粒的限制图谱(见于 Win95 平台)

位于窗口左侧的注释图标可加入额外的注释(线、圆、弧、箭头、文本等)。其他的格式或显示变化可通过查看菜单来实现，其中可选择性显示或隐藏序列标尺、网格线或酶标记。在查看菜单下也能以表格形式显示限制消化的完全概要，包括酶的名称、切点频率、酶切位置和识别序列。在帮助菜单下，用户可查看完整的 GeneTool 酶库，其中含有以字母顺序排列的酶名称和识别序列的列表。

6.4.6 版面编辑器

版面编辑器允许用户得以创建包含复杂文字的版面或文本图形(图 6.13)。这些 DNA 序列数据的复杂文字性(textual)描述常见于已发表的稿件中，通常需要在字处理器上进行冗长乏味的加工过程。为降低创建这种文本图形的难度，BioTools 开发了一个特殊的版面编辑器以加速和简化编辑过程。如图 6.13 所示，这个编辑器类似于 GeneTool 的序列编辑器[缺少 Feature Legend(特征图例)框]，但它增加了一些控制功能以调整输出。通过选择(利用鼠标)DNA 序列段来进行格式化，再点击 Grouping(集合)、CAPITALIZATION(大写)、Double Stranded(双链)、Translation(翻译)或 Show Restriction Sites(显示限制位点)等按钮，可改变或注释高亮化的序列。翻译按钮允许多框架翻译(1、3 或 6 个)，可采用单字母的 IUPAC 氨基酸密码或三字母密码。类似的是，Restriction Digest(限制消化)按钮允许利用文字(textually)对限制酶切位点位置进行注释描述，它采用与限制图谱视窗相同的对话框和选择程序。

通过 GeneTool 的打印预览器可实现其他格式化和注释选项。这个特殊的窗口便于用户查看文本，就像打印的结果一样，并能在 PostScript 图像的选择区域添加或改变文本、线、箭头、框或其他有用的注释。

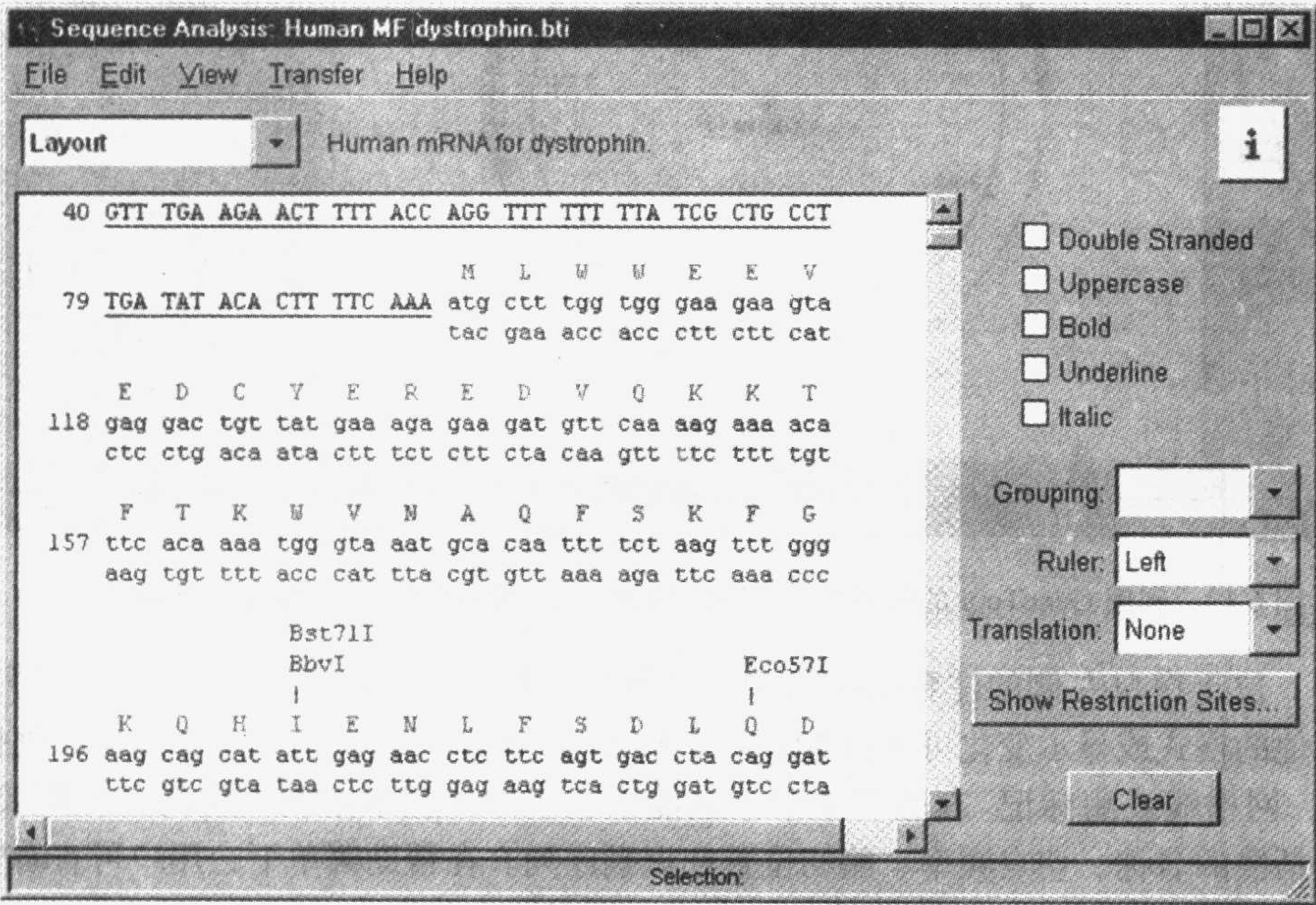


图 6.13 由 GeneTool 的版面编辑器产生的文字图形的例子(见于 Win95 平台)

6.5 基本程序特征——网络并行

PepTool 和 GeneTool 都提供一种叫做网络并行的独特的加速特征。网络并行允许用户同时在多个网络计算机上运行一个程序或过程。与在一个计算机上相比，在多个计算机上运行一个程序，其优势在于程序运行可被加速，其所耗时间会除以一个因子，相当于所用计算机的数目。与购买数百万美元的超级计算机相比，网络并行是一个更廉价的选择。事实上，考虑到许多实验室、大学和私人公司已拥有由多台个人计算机构成的网络，利用网络并行意味着可相对容易地免费获得10 个或 100 个网络计算机的功能(而不干扰网络中其他用户的操作)。BioTools 已在大多数苛刻的和耗时的数据库搜寻工作(Needleman-Wunsch 算法)中实现了网络并行(利用 PVM)。并计划把这个十分有用的功能延伸到其他耗时的操作中，如多序列比对(multiple alignment)、重叠群组装、二级结构预测和外显子鉴定。

6.5.1 数据库压缩

蛋白质和基因序列数据库比硬盘容量增加的快得多。GenBank 在 1998 年 4 月的发行(最后一次发行 CD)需要 12 个 CD 来装下所有的序列数据。幸运的是，通过互联网进入 NCBI 或 EBI 的 BLAST 服务器，使许多研究者快速进入这些庞大的数据库而不必寻找一个地方储存大于 10G 的数据或每次必须阅读一打 CD。但这些公共服务器在能进行的搜寻的形式和数据保存、发表、下载的方式上有一些限制。此外，不断增长的大学研究者和私人公司越来越关心查询公共数据库时发生的互联网安全性和防火墙缺陷。问题是：如何允许灵活地进入数据库并保持安全性，同时不必受每 6 个月购买一个新硬盘或每周购买一个新 CD 的烦恼。

一个答案是采用数据压缩技术。BioTools 利用以下事实：大多数生物数据只使用有限的字母，DNA 采用 4 个而蛋白质采用 20 个。这意味着 ASCII 字符集的大小可从每个字节 8bit 下降为 DNA 序列数据的大约 2.3bit 和蛋白质序列数据的 5bit。而且通过从数据库文本文件中去除空白，空的空格或多余的信息，并用特殊字符取代常规字符，就能在没有明显的信息损失的情况下获得更多的压缩。最后，通过把具有二重入口的多个数据库(如蛋白质序列)结合为一个非冗余数据库，将节省更多的空间。利用这些和其他压缩技术，BioTools 使蛋白质数据库的大小从 300M 下降到 60M，使 GenBank 数据库从 12G 下降到 3.2G。这意味着完整的数据库可发表在 2 CD 上(而不是 18 CD)和轻易地储存在常规的 4G 硬盘上。

尽管保持本地数据库与进入远程数据库相比更方便、灵活和安全，研究者将继续需要定期进入 NCBI 或 EBI 的高速设备和高度完整的数据库特征。为保持这个重要的数据库进入路线，BioTools 在它的 GeneTool 软件包中整合了 NCBI 服务器的万维网入口。

6.6 概要

尽管我们无法讨论 PepTool 和 GeneTool 的所有功能，从这个简短的综述中可以发现两个软件包在功能和易用性方面都十分优秀。此外，许多有用的革新，包括非平台依赖性 GUI 设计、网络并行、直接互联网连接、数据库压缩和多种增强或改进的算法，使得这两个程序在快速变化的生物序列分析领域显得特别有用。有关 PepTool 和 GeneTool 程序、算法及操作等更复杂的描述可从 BioTools 的网页上(www.biotools.com)获得，它们在相关的程序用户手册和在线帮助页中。

致谢

作者要感谢 Scott Fortin、Ann Leins 和 Debby Waldman 对稿件有益的批评指正。我们也要感谢 BioTools 的同事帮助我们制作了大量的插图。G. H. Van D. 得到了PMAC-MRC奖学金的资助，P. S. 得到了NSERC研究生奖学金和来自 Alberta 遗产基金的医学研究奖学津贴的资助。

(吴东林　译)

参 考 文 献

[1] Wishart, D. S., Boyko, R. F., Willard, L., Richards, F. M., and Sykes, B. D. (1994) SEQSEE: a comprehensive program suite for protein sequence analysis. *Comput.Applic. Biosci.* (now *Bioinformatics*) **10,** 121-132.

[2] Wishart, D. S., Boyko, R. F., and Sykes, B. D. (1994) Constrained multiple sequence alignment using XALIGN. *Comput. Applic. Biosci.* (now *Bioinformatics*) **10,** 687-688.

[3] Wishart, D. S., Fortin, S., Woloschuk, D. R., Wong, W., Rosborough, T., Van Domselaar, G., Schaeffer, J., and Szafron, D. (1997) A platform-independent graphical user interface for SEQSEE and XALIGN. *Comput. Applic. Biosci.* (now *Bioinformatics*) **13,** 561-562.

[4] Needleman, S. B. and Wunsch, C. D. (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* **48**, 443-453.

[5] Cattell, K., Koop, B., Olafson, R. S., Fellows, M., Bailey, I., Olafson, R. W., and Upton, C. (1996) Approaches to detection of distantly related proteins by database searches. *BioTechniques* **21,** 1118-1122.

[6] Upton, C., Mossman, K., and McFadden, G. (1992) Encoding of a homolog of the IFN-γ receptor by myxoma virus. *Science* **258,** 1369-1372.

[7] Upton, C., Stuart, D. T., and McFadden, G. (1993) Identification of a poxvirus gene encoding a uracil DNA glycosylase. *Proc. Natl. Acad. Sci. USA* **90,**4518-4522.

[8] Klein, P., Kanehisa, M., and DeLisi, C. (1985) The detection and classification of membrane-spanning proteins. *Biochim. Biophys. Acta* **815,** 468-476.

[9] Burset, M. and Guigo, R. (1996) Evaluation of gene structure prediction programs. *Genomics* **34,** 353-367.

7 Staden 软件包，1998

Rodger Staden Kathryn F. Beal James K. Bonfield

7.1 引言

几年来，我们一直致力于建立大规模测序项目的管理分析方法，所开发的软件在很多大型实验室和基因组中心使用。工作期间，我们设计了功能非常强大的图形用户界面。在此界面上，可以使用我们的序列组装和编辑程序 GAP4[1]。近来我们开始编写旧分析程序 NIP[2]和 SIP[3]的替代者，全部新程序都使用与 GAP4 相同的界面。旧程序在本书以前的版本[4]中进行了介绍，且大部分都没有改变，仍含在我们的软件包中销售。这里，我们概述我们的测序方法及新分析程序，所有这些都记录在我们 500 页的手册中，该手册可供打印或以超文本语言(HTML)文件的形式得到(http://www.mrc-lmb.cam.ac.uk/pubseq)。该站点内也含本章中所用的彩色图和获得软件包的信息。

7.2 管理测序项目的方法

7.2.1 引言

尽管很多工作已经自动化了，但是对于比较困难的重叠群的连接和编辑的决定等测序项目中的一些工作仍然需要人的判断。在我们的软件中，所有的过程包括序列组装都是全自动的，且我们提供的工具自动分析重叠群，提出帮助解决问题和连接重叠群的实验和模板。手动检查和编辑重叠群遵循受用户的注意力影响最小化原则，仅当保守碱基没有测到所需的精确水平时才受用户影响。这可使用 GAP4 的功能强大的重叠群编辑器完成。除了有自己的运算法则外，GAP4 还提供组装引擎 CAP2[5]、FAK1[6]及 PHRAP[7]的图形用户界面(见 7.4 节)。我们发明了储存来源于基于荧光的测序仪的跟踪数据和准确性值的 SCF 文件格式[8]，在最近的形式中(J. K. Bonfield 和 R. Staden，未发表资料)，储存来源于 ABI 测序仪的数据所需空间下降到十分之一，且可以在程序进行过程中转换相关的序列数据为所用的实验文件格式[9]。

7.2.2 预组装过程

组装前，从自动测序仪产生的序列，必须经过几个过程读出，通常包括转换为 SCF 格式、计算读出的碱基的准确性或置信值、按质截取、测序载体截取、克隆载体(如黏粒)去除、重复序列加上标签。用我们的软件包，这些步骤的每一步都被一个独立的特定程序(图 7.1)完成，而整个过程则被程序 PREGAP[9](见 7.4 节)控制，此程序能在单次运行中控制任意次阅读。

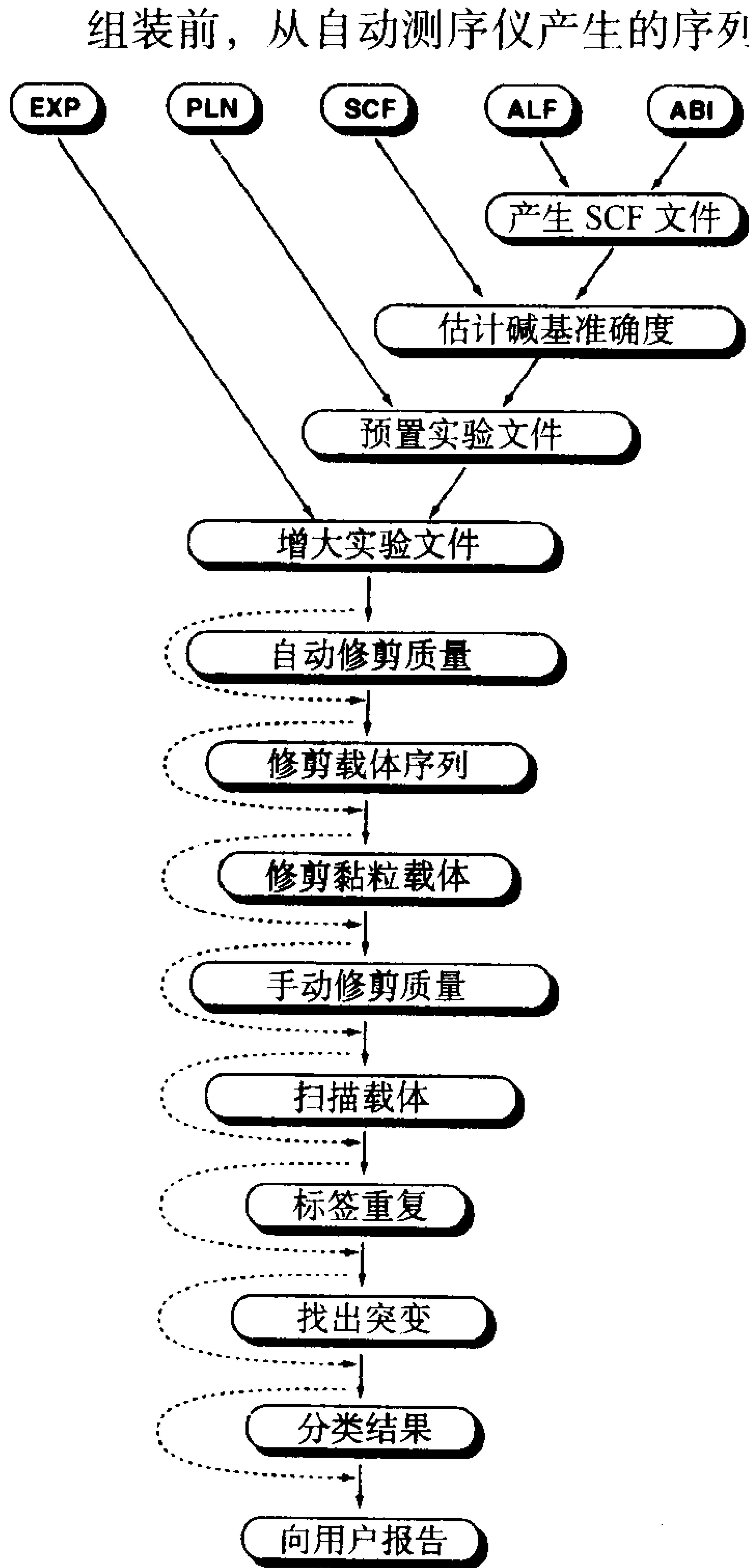


图 7.1 PREGAP 处理步骤

输入到 PREGAP 的是一个文件，它含有所有要处理的测序仪文件的名称。输出通常是一个 SCF 文件，及每个处理的输入文件的实验文件，加上一个文件名的新文件，它含有所有新产生的实验文件名称，准备传给 GAP4。也就是说，PREGAP 为每一次阅读形成一个起始实验文件，然后按所需的处理步骤依次传送。注意，使用外部组装引擎(如 PHRAP)时，必须使用文件名来编码它们所包含的有关碱基判读的数据，而使用 PREGAP 和实验文件则不需要。PREGAP 是模块化的，并且很灵活，能根据实际工作需要取舍。它能设置成全自动工作，也可以设置成部分交互式运行。

7.2.3 GAP4 简介

GAP4 提供全套装配方法，检查安装情况，通过重叠群端的可读数据和/或低质量数据找出重叠群间的连接，建议补加特殊测序实验完成序列连接或克服数据的其他缺陷，检查共有序列的准确性及将重叠群编辑到所需水平的可信度。正如所有测序者所知，如果试图处理两次碱基判读的每个差异，需要花很长时间，而且正是低质量的数据引发了组装问题。因此，我们支持且采用[10]在 GAP4 内的共有算法中使用碱基准确性估计，并把这些值考虑在内。在这种方法中，在质量好的数据与质量差的数据间存在差异时的地方，只需共有计算，而不必进行编辑，

用阈值来确保序列达到所需水平的可信度，确定 A、C、G 或 T。如果数据的可信度达不到要求，则使用“-”或“N”。由于从头到尾，整个 GAP4 的共有计算是做出决定的基础，有任何变化都能通过更新反映出来，因此，用户只需要关注碱基不是 A、C、G 和 T 的共有位点。注意，我们刚刚描述的是一种理想状态，在此存在可信碱基准确性估计，但是我们不考虑用我们的程序(EBA)计算的每个值都充分可信(EBA 简单地提供一个让我们建立相关方法的值，见 7.4 节)。我们意识到程序组能产生准确度值(或可信值)，且希望马上可以得到可靠的方法。

用户界面使用户能以恰当的方式显示和操作数据，它在帮助用户处理困难问题时起重要作用，且简化并加速测序项目。图 7.1~图 7.6 中显示的图形包括 Contig Selector、Contig Comparator、Template Display、Quality Plot、Restriction Enzyme Map 及 Stop Codon Map。使用 Contig Editor 和 Join Editor 编辑配对的碱基判读结果，这些判读中的每一个，用编辑光标能够显示及滚动判读的踪迹数据。

GAP4 的显示互相联络，并通过指针相链接。例如，如果 Contig Editor 在一个重叠群中使用，重叠群也在 Template Display 窗口中显示，Contig Editor 的指针位置也显示在 Template Display 中。相似地，在 Template Display 中，用户能用鼠标拖拉 Contig Editor 的编辑指针。而且，如果 Contig Editor 下显示踪迹，它们也将在 Template Display 控制下滚屏。同一重叠群的任何数可以同时显示，包括同类型的几个显示，如几个 Contig Editor。

7.2.3.1　GAP4 文本和图形窗口

GAP4 的主窗口的例子(其中,除主菜单外,与 SIP4 和 NIP4 相同)如图 7.2 所示。

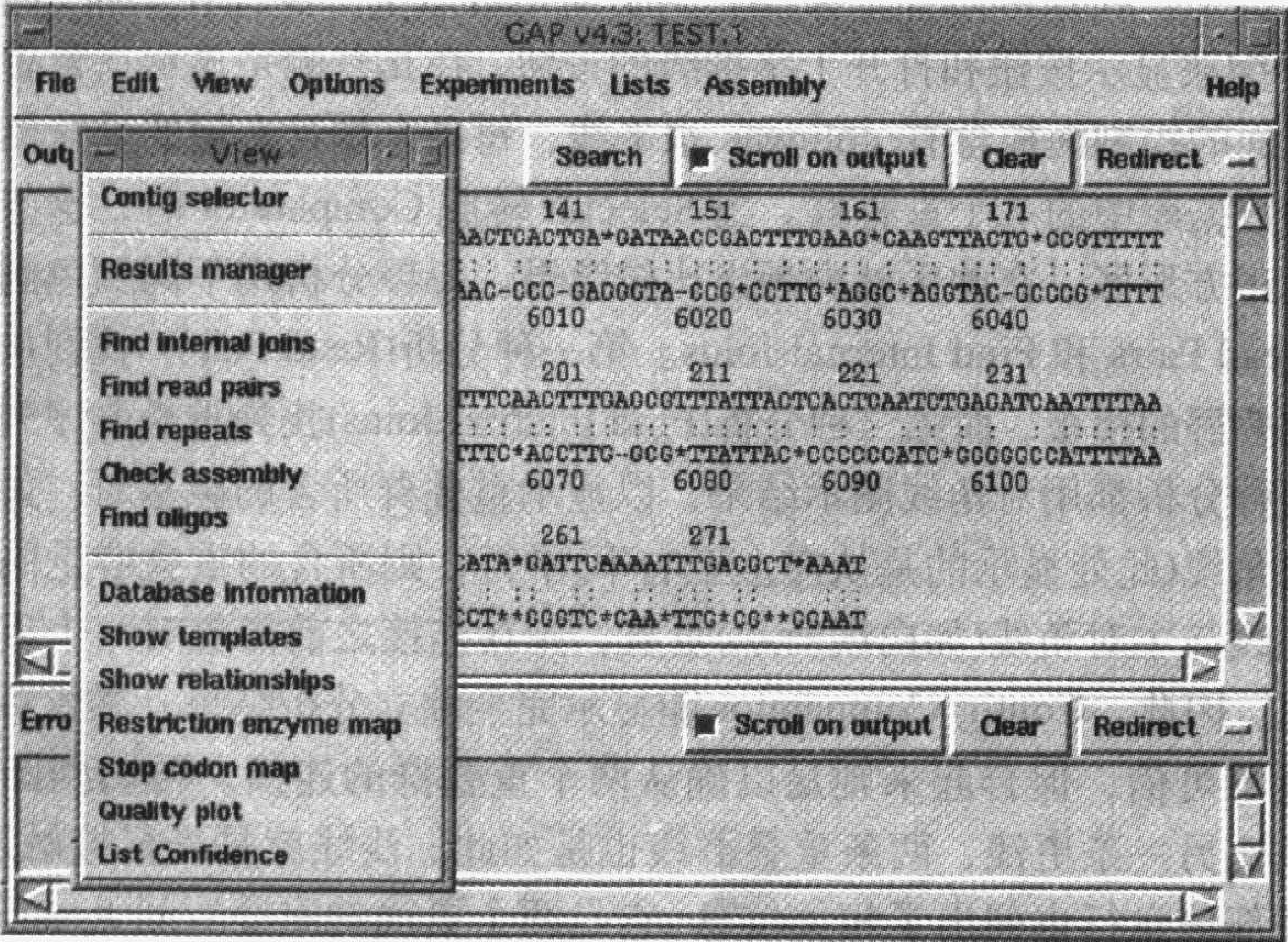


图 7.2　GAP4 主窗口

它含一个接收文本结果的输出窗口及下面的显示错误信息的区域。在窗口的顶部是一行菜单：File 菜单含数据库打开和拷贝功能及保守序列文件产生的例行程序。其他菜单的大多数重要项目见表 7.1。

表 7.1　GAP4 的菜单特性

Edit 菜单	View 菜单	Experiment 菜单	Assembly 菜单
Edit Contig	Contig Selector	Suggest Primers	Normal Shotgun
Order Contigs	Results Manager	Suggest Long Reads	Directed Assembly
Join Contigs	Find Internal Joins	Compressions and Stops	Independent Assembly
Break Contig	Find Read Pairs		CAP2 Assembly
Disassemble Readings	Find Repeats		FAKII Assembly
Complement Contig	Show Templates		PHRAP Assembly
	Restriction Enzyme map		Screen Only
	Stop Codon map		
	Quality Plot		
	Check Assembly		

7.2.3.2　Contig Selector 和 Contig Comparator

一旦一个数据库被载入 GAP4，将以图形方式显示出 Contig Selector。当进行重叠群的比较分析时，Contig Selector 自动转变为 Contig Comparator。图 7.3 显示的是一个例子。在顶部是三个菜单，其下面是进入下一步操作的按钮(见 7.2.3.6 节)、用于缩放显示按钮和打开十字准线的按钮，在其右侧是显示在重叠群中位置的框，下面的控制面板含有 colinear 水平线，每条末端含有短的垂直棒，代表正在显示的数据库中的 7 个重叠群。当 Selector 转成 Comparator 时，这些线在右角复制产生正方形区，在此区域内画出比较结果。每种分析用不同颜色画出结果，如区分 Read Pairs 和 Find Internal Joins。第一种分析(Read Pairs) 找到从两端测序并跨越重叠群的模板，而第二种分析(Find Internal Joins)找到重叠群序列间的匹配序列。两种分析都用对角线图示结果，以显示重叠群所含的碱基对。如果画出的线与主对角线(此处显示的是从左上至右下)平行，则所含的重叠群是同向的，如果画出的线与主对角线(如白线所示)垂直，则在连接之前需要对一个重叠群形成互补序列。因此，Contig Comparator 能显示完全独立分析结果，尽管各个独立的结果可能不可信，但合起来则足以确认两个重叠群的连接。再回到图上，对角线的一条比另一条更浅，两条重叠群线也是如此。浅线就是光标接触过的结果，显示区底部信息行也列出了相关信息，它是两个更浅的重叠群间 Read Pair(读对)匹配。

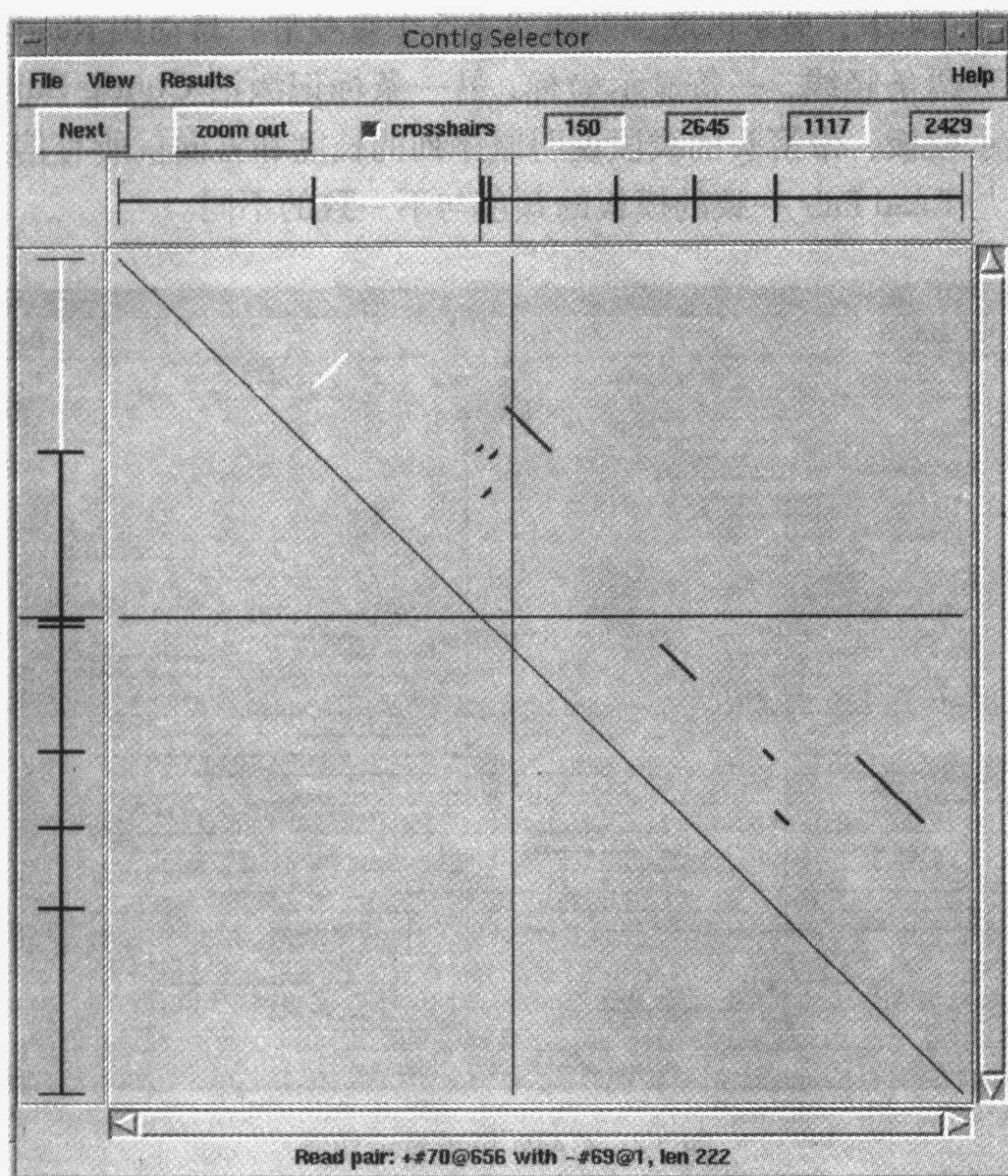


图 7.3 GAP4 Contig Comparator(GAP4 重叠群比较器)

在代表一个重叠群的一根线上点击一下鼠标键，用户能调出一下拉式菜单，菜单含对重叠群操作的列表。与此相似，在图形结果上点击，含相关操作列表的菜单显示出来。例如，如果用户点击 Find Internal Joins(找内部连接)结果，Join Editor(连接编辑器)对两个重叠群开始在报告的匹配位上比对。Join Editor 等于两个 Contig Editor(重叠群编辑器)(下文中介绍)，而且显示出两个比对重叠群的差异，且一旦编辑完成就能将它们连接。

7.2.3.3 Template Display(模板显示)功能

如图 7.4 所示，Template Display 功能为重叠群组或单一重叠群提供了图解。所能显示的信息有：读数、模板、标签、限制酶位点、标尺、共有性程度。与 Contig Selector 一样，如果鼠标移过显示中的任何项目，这些项目的文本数据信息将在底部的信息行中出现。在鼠标的控制下，重叠群和重叠群组的坐标位置连续显示在窗口顶部的两个盒子里。拖拉显示区中代表水平位置的线条能改变重叠群的顺序，并且所有相关的数据都重画(这也可由 Contig Comparator 完成)。含 5 个重叠群的

例子显示在图 7.4 中，显示区底部的线代表 5 个重叠群，自动用 Read Pair 数据隔开。还可见到两条竖线，一条显示坐标，另一条标记激活 Contig Editor 的位置。大量横线表示模板，带箭头的线段说明原序列的长度和方向。所有信息按颜色分类，例如，与 Read Pair 一致的模板的颜色与不一致的不同。

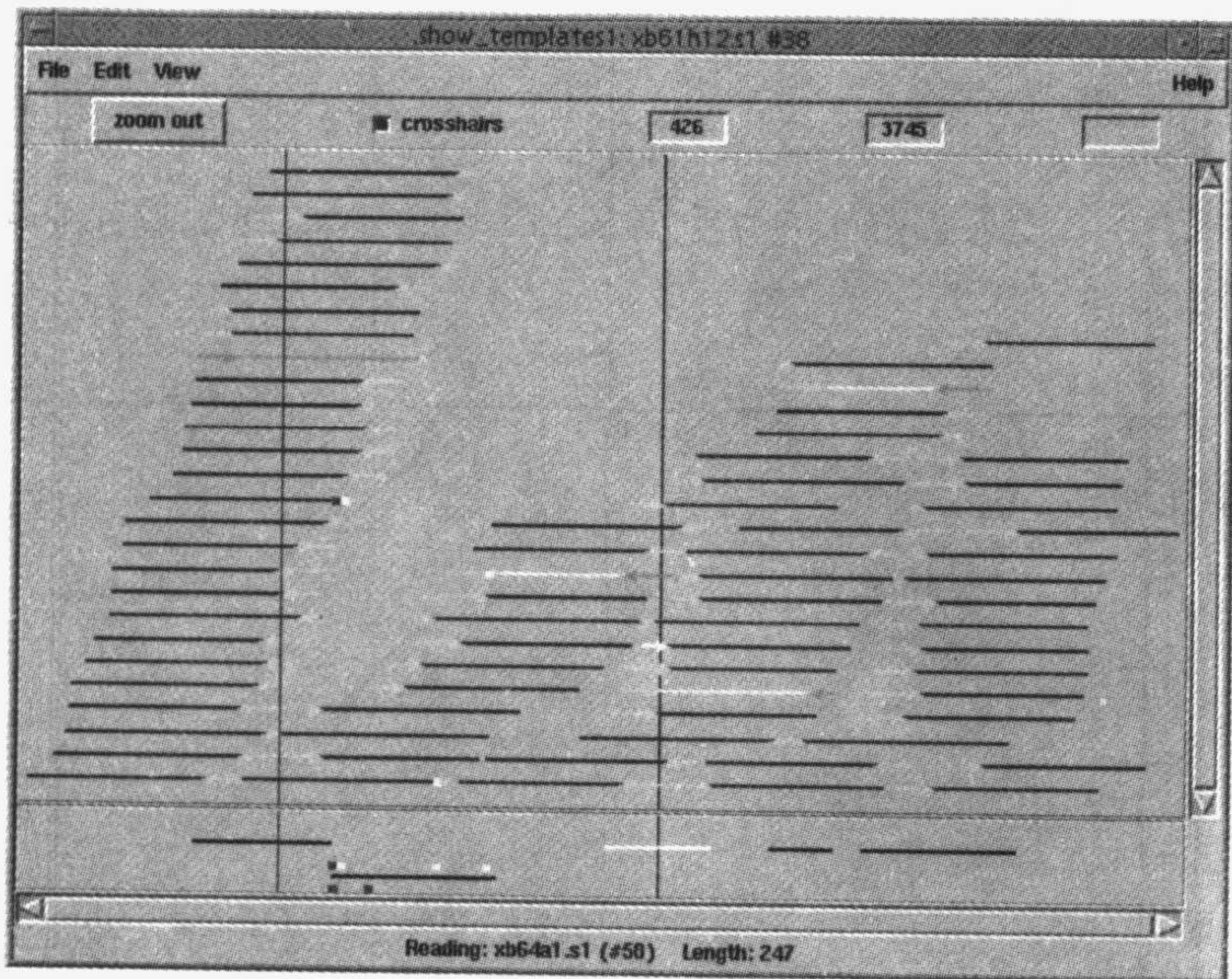


图 7.4 GAP4 模板显示功能

7.2.3.4 Contig Editor(重叠群编辑器)

重叠群编辑器(图 7.5)是 GAP4 最重要的组件之一。它含有优秀的检索功能，定位和处理重叠群问题耗时最少。在实际的应用中，读数时，它可以得到碱基可靠性估计或置信度值，能搜索共有序列，筛查出达不到所确定的精确水平的位置。这种共有性在线计算出来，并在读出的序列改变时更新。当得不到置信度值时，使用在线共有性计算功能(每条链分开处理，保证独立地确定两条链序列)，差异就引导用户的注意力到需要关注的地方。另一种搜索将所有编辑位置定位到在读出的(涵盖位点的)源序列中不出现共有碱基的地方，因此，能检查编辑是否正确。通常，使用搜索功能时，程序设置成自动显示引起问题的读序踪迹，也显示每条链的好的读序。

在读序和共有序列中的每个字母的可信度值用灰度程度显示。各个编辑位置上的颜色则说明所用的编辑方法的种类。xb62c2. s1 读序中的黑段区是名为标签(tag)的标记。该程序根据不同目的使用一套不同标签，每一型有一独特的颜色，每个标签段有一个相关的可编辑的文字串用于记录注释。在所有 GAP4 的显示中，标签都可以变成可见的。

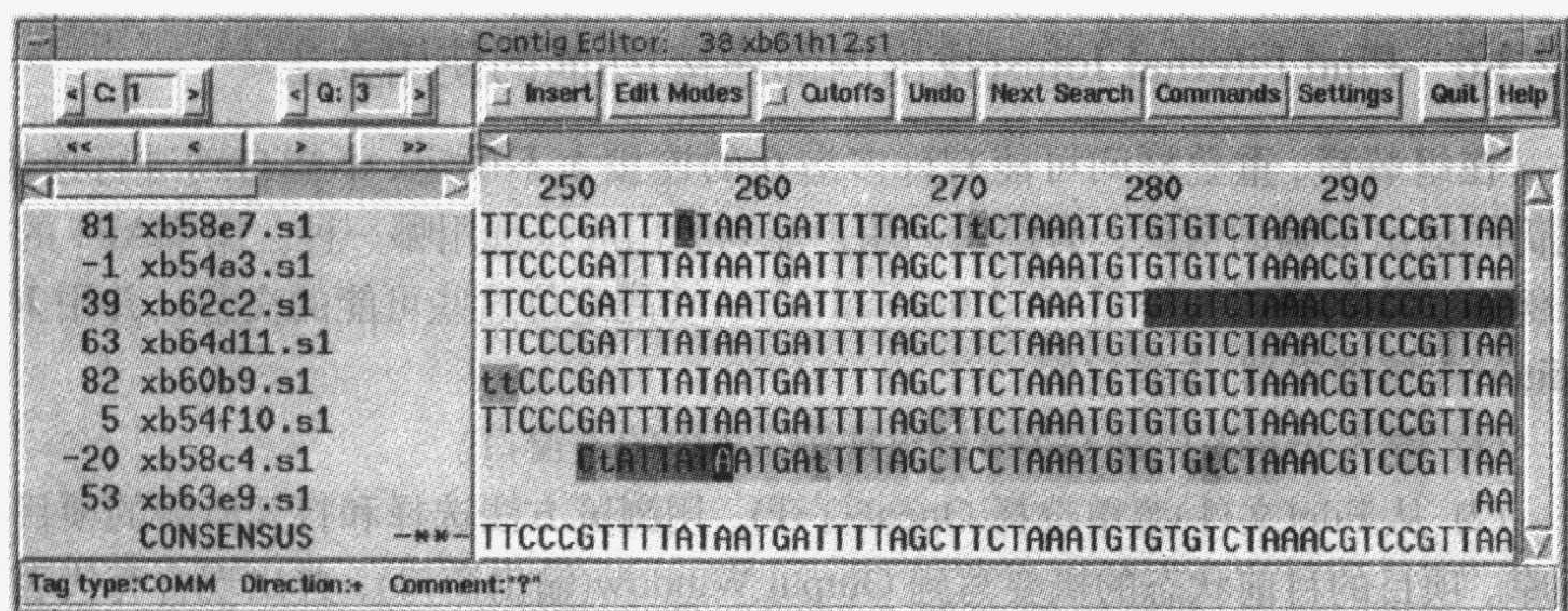


图 7.5　GAP4 重叠群编辑器

7.2.3.5　GAP4 方法

本节中，我们概述 GAP4 辅助测序项目的功能。正如 7.2.1 节中提到的，GAP4 含有几个我们自己加上的 CAP2[5]、FAKII[6]和 PHRAP[7]等全局性方法的组装算法。为了加工大批鸟枪法数据，这些数据按完全覆盖所测序列的目的收集，使用全局性算法最好，特别是如果序列含有长的重复序列更是如此。此外，当得到为解决问题和填充间隔的新读序时，我们推荐使用我们自己的鸟枪法组装算法。对于读数的大概定位已知的测序项目，我们的定向组装算法是很有用的。需要注意 GAP4 数据库储存有大量额外信息，而不仅有读序。主要的附加数据提供给 CAP2、FAKII 和 PHRAP，用于读写实验文件，因此经过 GAP4 的处理，数据中有用的信息都能发挥作用。

我们使用这些额外信息的一种方式是借此检查来自同一测序模板的读序的相对顺序和方向。这种方式用于在组装过程中找到可能的错误，以及找出最可能的从左到右的顺序和重叠群的方向。Order Contigs 程序例行计算最可能的顺序，单击结果，用户能调用 Template Display(图 7.4)。

有时，进入 GAP4 数据库中的数据可能不正确(如模板名可能错误)，因此，在重叠群排序过程中可能会产生错误。如果出现这种情况，在 Template Display(模板显示区)内用鼠标移动重叠群到适当的位置上可以改变顺序。如图 7.3 所示，这类 Read Pair(读碱基对)信息也能显示在 Contig Comparator(重叠群比较器)中，这在结合 Find Internal Joins(找出内部连接)功能时尤其有用，Find Internal Joins 功能可以比较重叠群末端，寻找重叠群间可能的匹配。

采用与 Contig Editor(重叠群编辑器)相同的共有序列算法，程序能得出最终成品序列。其他功能可以分析重叠群，找出不足以确定的区域，或能帮助连接重叠群。这些常规方法是实验建议方法，可以产生模板名和实验类型列表，其结果以易于解析的格式写到磁盘上。

7.2.3.6 Find Internal Joins(找出内部连接)功能的使用方法

在组装后，重叠群间可能有许多潜在的连接太不确定，不能自动进行连接，但仍能给出正确的信息，有经验的用户能判断出应该选用哪一种。用 GAP4 的一个例子，我们说明怎样用 Find Internal Joins 功能定位这些可能的交叠，并用 Join Editor(连接编辑器)连接它们。

(1) 键入 gap4&启动 GAP4，出现图 7.2 中的主窗口。

(2) 从 File(文件)菜单选择 Open(打开)，用浏览方法选择和打开所需的项目数据库。项目的目前状态的概要写到 Output Window(输出窗口)，并出现含有代表所有重叠群的线条的 Contig Selector(重叠群选择器)。

(3) 从 View(视图)菜单选择 Find Internal Joins，出现图 7.6 所示的对话框。

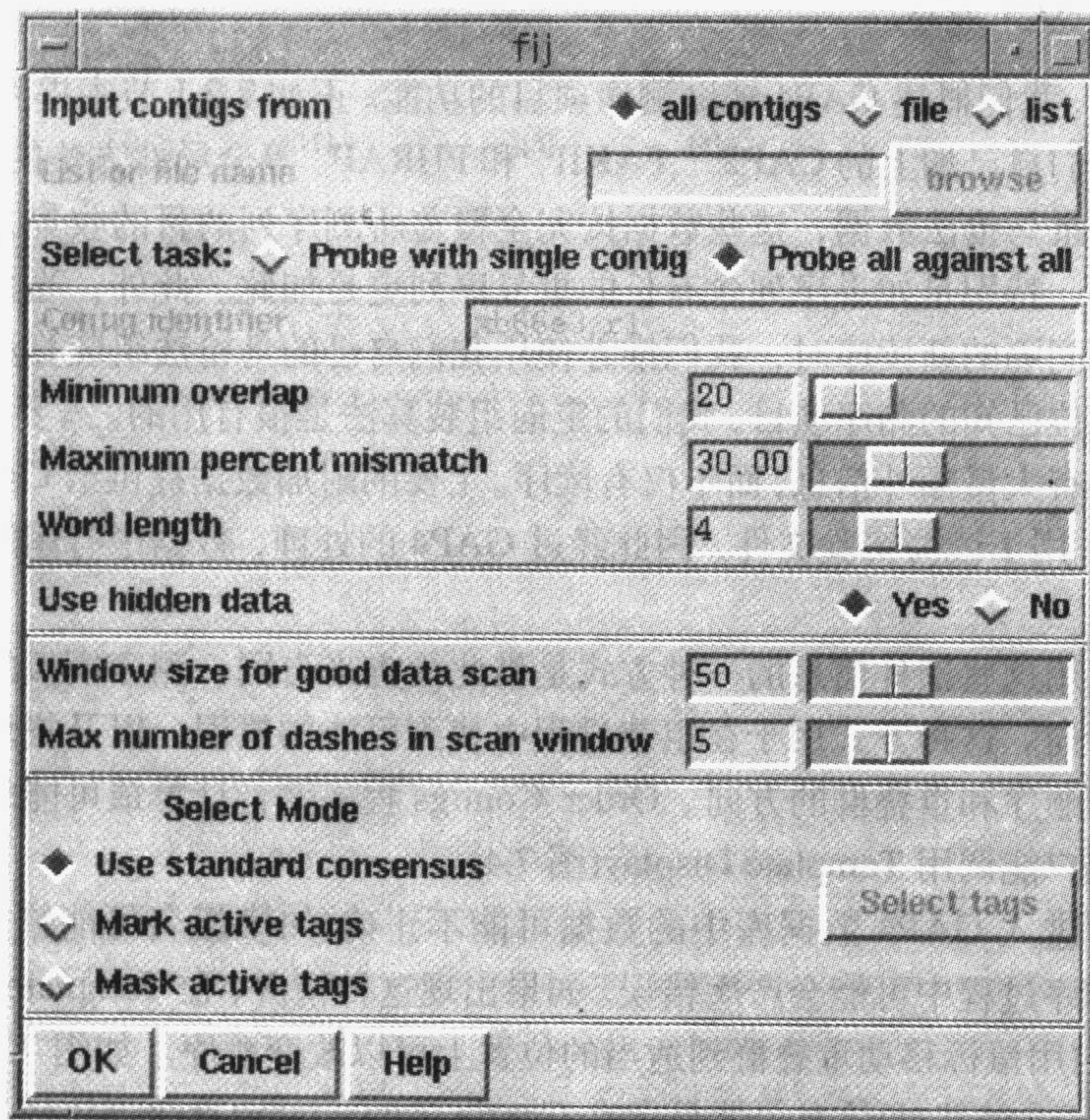


图 7.6 GAP4 的 Find Internal Joins 程序对话框

(4) 默认的参数设为比较每个重叠群、双向进行、互为对照、两个重叠群间需最少交叠 20 个碱基、在比对后最大错配百分数为 30%、字长为 4。如果必要可以在文本窗口键入新值或用相邻的滑条改变设置。

(5) 为了扩展含最低质量数据末端的重叠群，选择“Yes”键激活 Use Hidden data(使用隐藏数据)，将 Window size for good data scan(好数据筛查窗口大小)值和

Max number of dashes in scan window(筛查窗口最大未知点数)值分别设为 50 和 5。这意味着加到重叠群末端的数据不含超过 5 个未知碱基("-"或"N")的长度 50 的片段。

(6) 点击 Mask active tags(屏蔽激活的标签)。此项功能意味着选定的标签类型覆盖的序列片段不用于初始匹配(但用于其他地方的初始比对)。例如，标上 Alu 重复序列的标签的片段如果与边界外的序列也有一个好的匹配，将仅用于可能的交叠。默认状态，标签类型 ALUS、REPT 和 MASK 将被屏蔽。要改变此设置，点击 Select tags(选择标签)，出现一个含所有类型标签名的对话框。点击相应的键对所需的类型进行取舍。点击 OK 退出标签选择窗口。

(7) 在 Find Internal Joins(找出内部连接)对话框中单击 OK，搜索将开始进行。搜索进行过程中，出现一个忙光标，但是搜索过程通常很快。如果窗口仍不出现，Contig Selector(重叠群选择器)窗口将自身转到 Contig Comparator(重叠群比较器)，准备接收产生的结果。结果将像 7.2.3.2 节所述的一样画出，交叠的比对也将写到输出窗口。

(8) 点击 Result(结果)菜单，向下拖到 Find Internal Joins 结果。显示一下拉菜单。向下拖到 Sort matches(分类匹配)按钮，然后释放，然后用 Use for Next(为下一个使用)按钮重复。这使 Find Internal Joins 的匹配成为错配的百分比的顺序，按 Next 命令，结束本步程序。

(9) 点击 Contig Comparator 窗口左上部的 Next 按钮，含来自于 Find Internal Joins 搜索的最佳比对的 Join Editor 窗口出现。

(10) 在 Join Editor 窗口中，点击 Align(比对)按钮，两个共有序列及其读序用填充符(*)比对排列。

(11) 用各种运动方法沿交叠的全长序列滚屏，观察任何不满意的踪迹(在编辑器窗口中的序列上双击可达到目的)。

(12) 点击 Quit(退出)按钮退出 Join Editor，如果确信连接是对的，在出现的对话框中点击 OK 按钮。画图自身通过连接两条重叠群线重排，其他结果也相应地重新定位。在 Template Displays 中也显示目前已连接的重叠群。

(13) 点击 Next 按钮重复运行过程，进行下一步最好的连接。继续此步，直到所有结果都检查过。任何时候，双击画在 Contig Comparator 中的任何 Find Internal Joins 结果，将调出一个 Join Editor。Next 按钮是按最优顺序运行的捷径。

7.2.3.7 更多的 GAP4 内容

显然，GAP4 含有很多其他功能和能力，如打断重叠群或移去读序。尽管在此我们没有强调使用程序的交互式操作，但这点仍值得关注，大多数 GAP4 的功能可以从脚本运行。由 James Bonfield 撰写的解释怎样用 GAP4 脚本语言写程序的手册可从我们的网站上得到，网站地址为：`http://www.mrc-lmb.cam.ac.nk/pubseq`。

7.3 分析成品序列的新程序

正如 7.1 节所述，我们已经开始建立一套新程序，用于分析成品序列，这受益于我们为 GAP4 设计的用户界面。在写程序的时候(1998 年 5 月)，我们把 SIP4(K. F. Beal，J. K. Bonfield 和 R. Staden，未发表)视为一个优秀的富有特色的序列比较程序，但是，NIP4(K. F. Beal，J. K. Bonfield 和 R. Staden，未发表)，我们的核酸分析程序，尽管有好的用户界面和一些有用的功能，但仍需要工作到作为实验室内实现此目的的主要程序的阶段。像我们对 GAP4 所做的一样，我们的计划是其他人易于加算法到 NIP4，提供从外挂程序输入结果的途径。两个程序都对序列数据库提供功能强大的易于使用的界面。

7.3.1 序列比较程序 SIP4

SIP4 代替 SIP 程序，SIP 程序是最初在名为 DIAGON[3]程序下发表，用于比较序列组，找出相似区。它有与 GAP4 相似的用户界面。但是它的主要图形窗口(图 7.7)是一个交互式点矩阵显示，用于显示和分析比较结果。该程序含有几个比较序列的方法，如 DNA 对 DNA、蛋白质对蛋白质或 DNA 对蛋白质，从很快但不特别敏感到很慢但敏感的程序都有。所有的比较算法，局部[11]和全局[12]比对方法都形成图形结果，比对结果能在序列水平上显示。

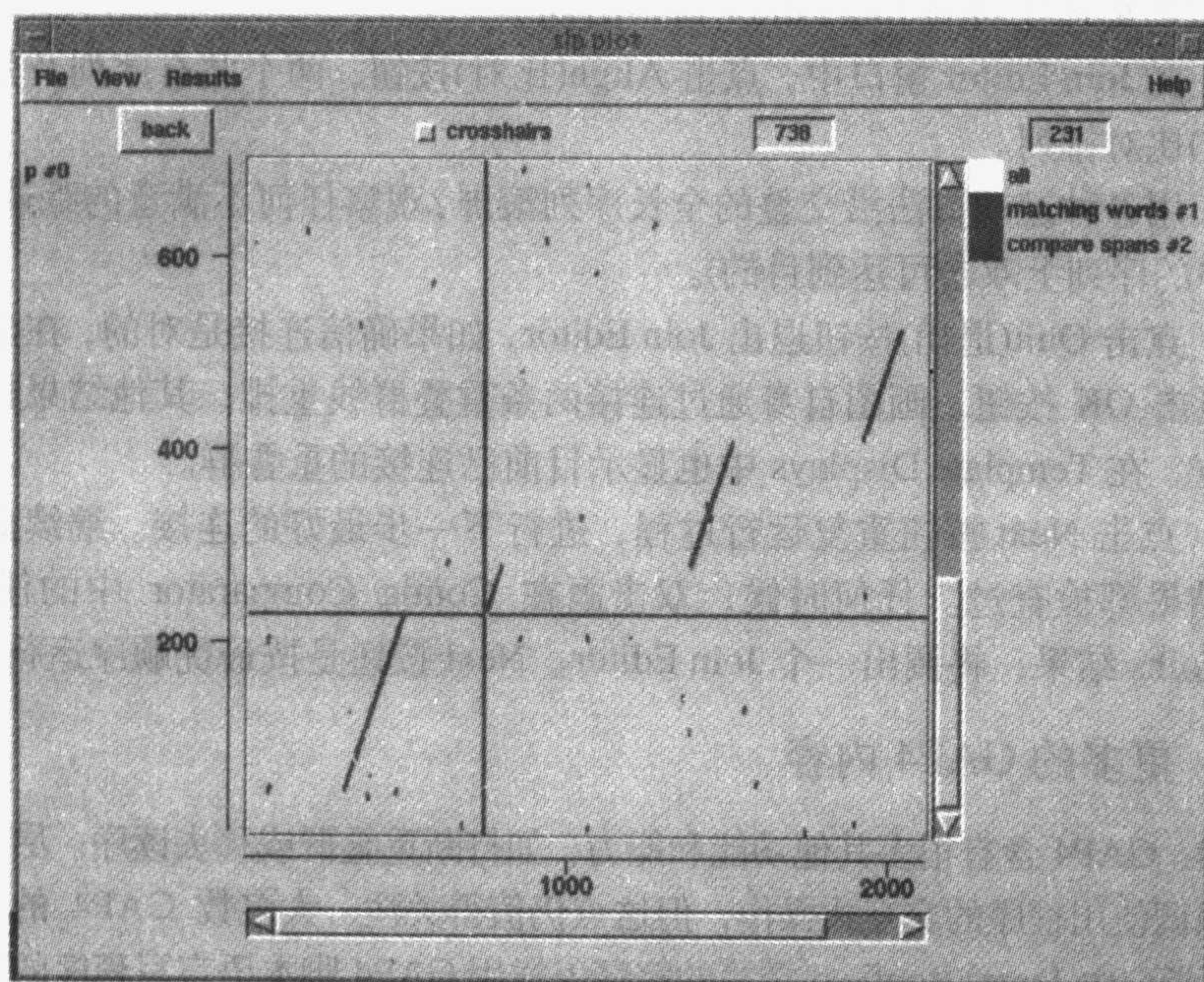


图 7.7 SIP4 程序的图形显示图

在图 7.7 中，我们显示了基因组 DNA 的一个区和其 mRNA 的比较结果。能用任意数量的比较方法和任意一套阈值，并且叠加到一个单点上，每一套结果用不同的颜色显示。此外，每套结果能移到自己的显示窗口，也可选择一个叠加上。图能被缩放及在 *X* 和 *Y* 平面上滚屏。其坐标显示在顶部的格子里。对任何结果，可滚动显示的序列比对窗口可以如图 7.8 所示显示：一个序列显示在上面，另一个序列显示在下面。每个都能独立滚动，如果 Lock(锁)按钮被按上，滚动时上下序列联动。按上 Nearest match(最相近的匹配)按钮将使序列滚动显示，直到最相近的匹配序列块出现在窗口的中心。此外，图形窗口中的坐标能用于滚动显示序列，以便能详细检查匹配。图 7.7 和图 7.8 对照显示的位置和序列位于内含子和外显子交界处。

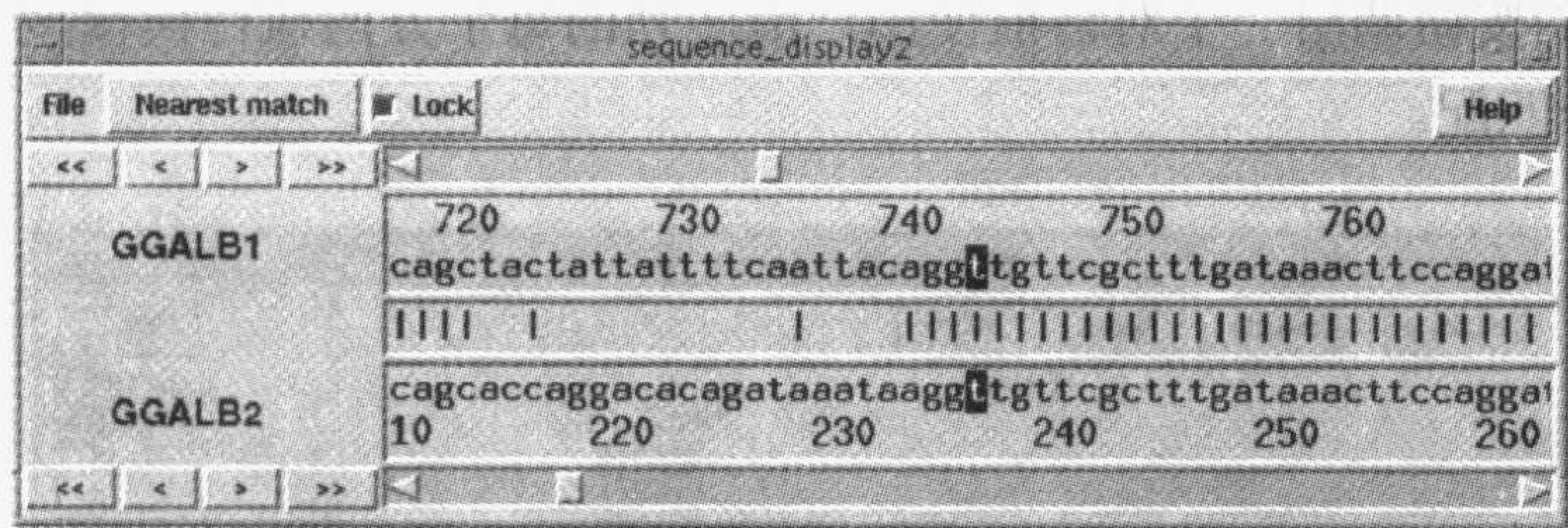


图 7.8　SIP4 程序的序列显示图

7.3.2　核酸序列分析程序 NIP4

NIP4 代替 NIP 程序，NIP 程序是最初在名为 ANALYSEQ[2]程序下发表，有与 SIP4 相似的用户界面。主要图形显示的例子如图 7.9 所示。我们的一种基因预

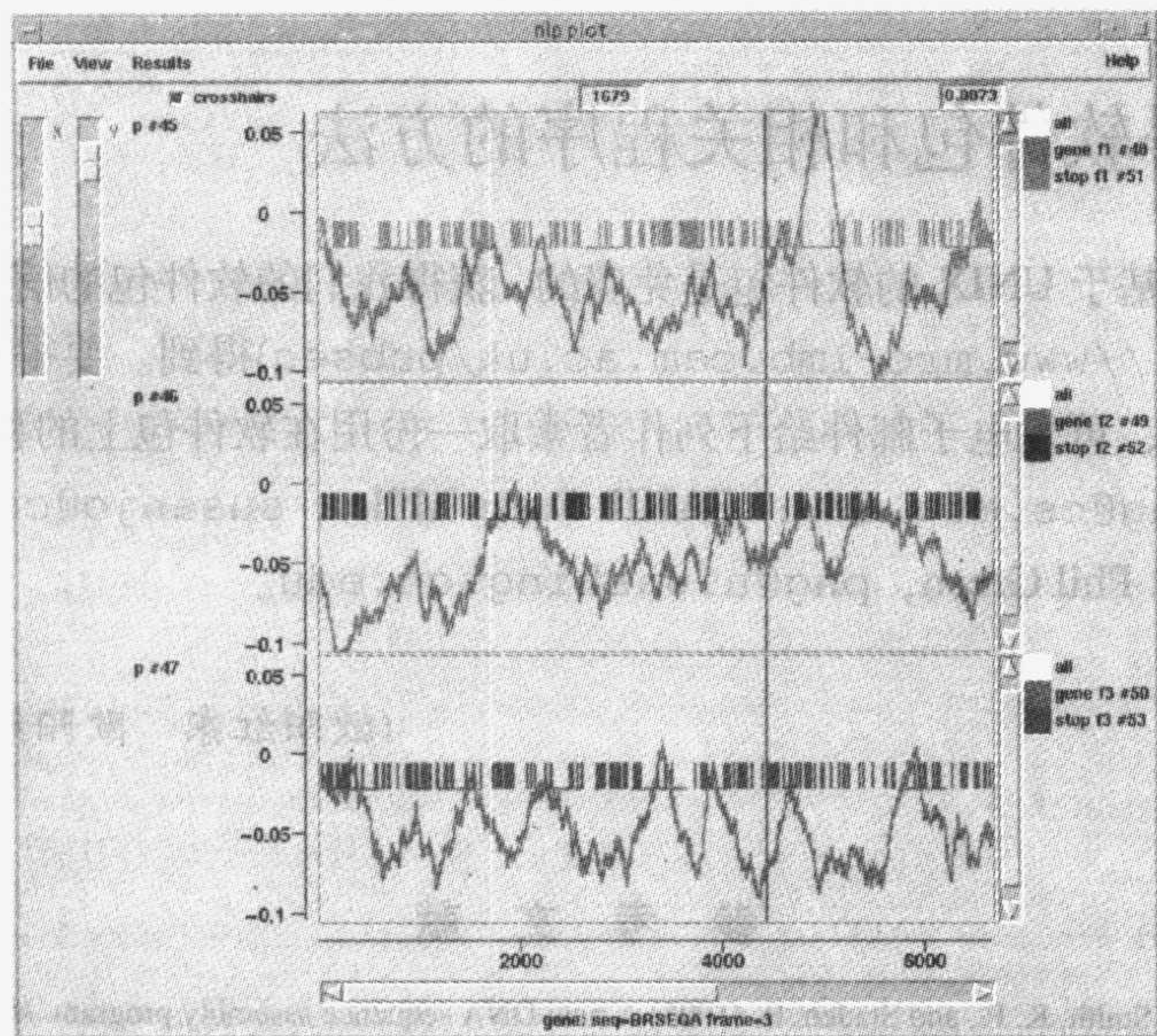


图 7.9　NIP4 程序的图形显示图

测方法画出3个读框中的每一个可能的编码，图上终止密码子被覆盖。每个结果(如一个读框的终止密码子或可能的一个读框的编码)用一种颜色画。用顶部左边的刻度条，图能被缩放及在 X 和 Y 平面上滚屏。数据视图也能通过图周围滚动条在 X 和 Y 方向移动。单个结果能被拾起放到新的位置上进行覆盖，也能放在独立的窗口。坐标线是图中的一对浅颜色的线，其坐标值在显示区顶上的格子内可见到。将图分成两部分的深色的竖线是 Sequence Display(序列显示)窗口的指针位置，图 7.10 所示的是它的一个例子。Sequence Display 窗口能用图形窗口中的坐标滚动，也可使用它自己内部的方法滚动。在展示的例子中，显示区显示序列的 2 条链、3 个翻译框、限制酶切位点。其中的每种显示都能用 Setting(设置)菜单激活或关闭。Search(搜索)按钮调出一个对话框使用户沿序列的每个方向执行模糊搜索，序列能滚动到每个匹配。

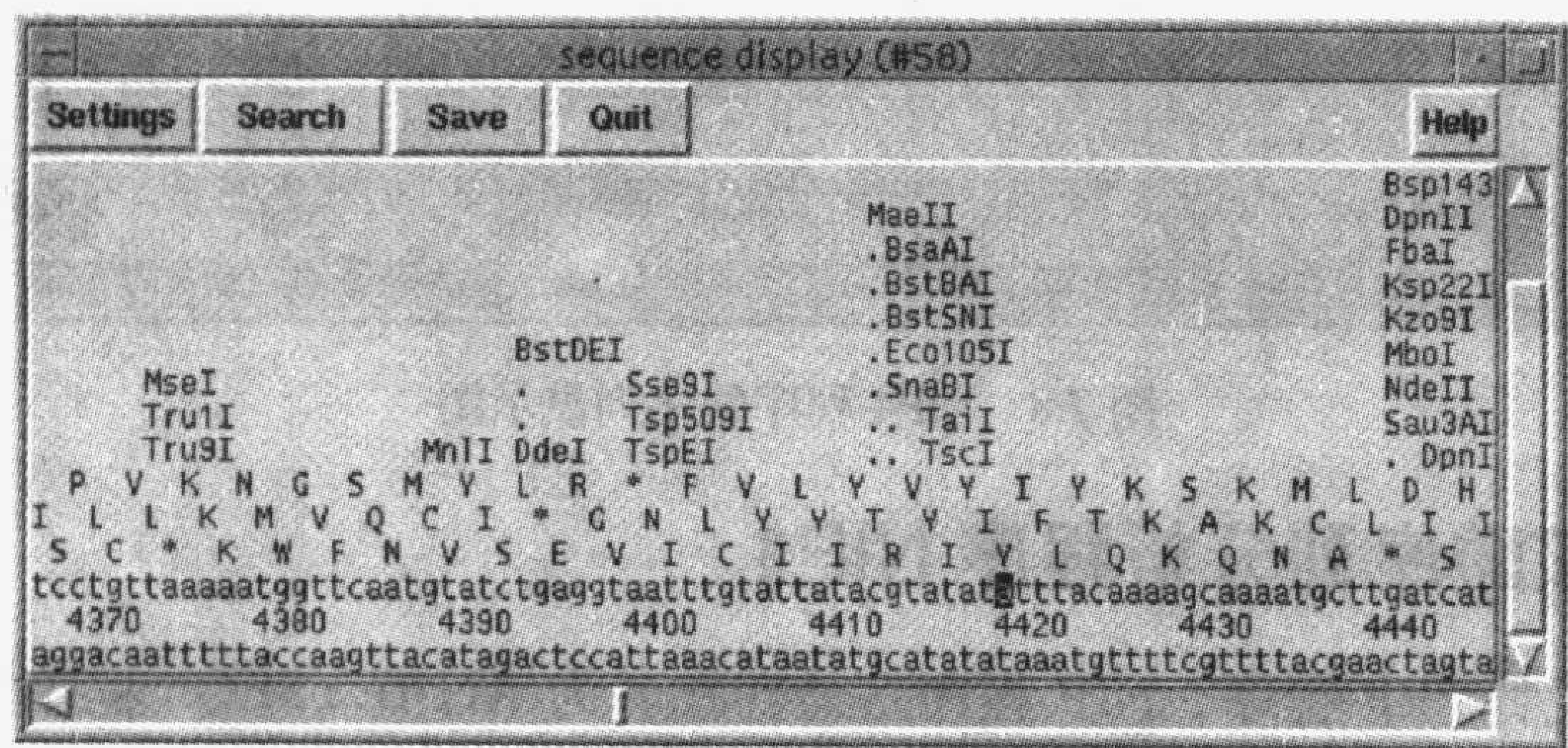


图 7.10　NIP4 程序的序列显示图

7.4　获得软件包和相关程序的方法

学术用的基于 UNIX 的软件包是免费的。获得我们的软件包的信息可以从万维网站点(http://www.mrc-lmb.cam.ac.uk/pubseq)得到。要得到能在 GAP4 内使用的程序，可发电子邮件给下列作者索取一份用在软件包上的程序。CAP2：黄小秋，huang@cs.mtu.edu；FAKII: Susan Miller，susanjo@cs.arizona.edu；PHRAP: Phil Green，phg@u.washington.edu。

(欧阳红东　欧阳红生　译)

参 考 文 献

[1] Bonfield, J. K., Smith, K. F., and Staden, R. (1995) A new DNA sequence assembly program. *Nucleic Acids Res.* **23,** 4992-4999.

[2] Staden, R. (1984) Graphic methods to determine the function of nucleic acid sequences. *Nucleic Acids Res.* **12,** 521-538.

[3] Staden, R. (1982) An interactive graphics program for comparing and aligning nucleic acid and amino acid sequences. *Nucleic Acids Res.* **10,** 2951-2961.

[4] Staden, R. (1994), in Methods in Molecular Biology, vol. 25, *Computer Analysis of Sequence Data, Part II.* (Griffin, A. M. and Griffin, H. G., eds.) Humana Press, Totawa, NJ, pp. 9-170.

[5] Huang, X. (1996). An improved sequence assembly program. *Genomics* **33,** 21-31.

[6] Myers, E. W., Jain, M., and Larson, S. (1996) Internal report, University of Arizona.

[7] Green, P. H. (1997) Pers. comm.

[8] Dear, S. and Staden, R. (1992) A standard file format for data from DNA sequencing instruments. *DNA Sequence* **3,** 107-110.

[9] Bonfield, J. K. and Staden, R. (1996) Experiment files and their application during large-scale sequencing projects. *DNA Sequence* **6,** 109-117.

[10] Bonfield, J. K. and Staden, R. (1995) The application of numerical estimates of base calling accuracy to DNA sequencing projects. *Nucleic Acids Res.* **23,** 1406-1410.

[11] Huang, X. Q. and Miller, W. (1991) A time-efficient, linear-space local similarity algorithm. *Adv. Appl. Math.* **12,** 337-357.

[12] Huang, X. Q. (1994) On global sequence alignment. *Comp. Appl. Biosi.* (now *Bioinformatics*) **10,** 227-235.

8 利用免费软件建立多用户序列分析系统

Brian Fristensky

8.1 引言

虽然分析分子序列的商业软件包很多，但普遍价格昂贵。也有许多基于网页应用的平台能用于序列分析，但在远程服务器上，且网页界面不能储存数据，因而使用起来也不方便。一个好的选择就是在本地服务器上建立序列分析系统。BIRCH(生物学研究计算机分级结构)就是这种系统的一个例子(http://home.cc.umanitoba.ca/~psgendb 和参考文献[1])。BIRCH 作为一种工作平台，它比较成熟，含各种序列分析工具和软件，这些软件在执行任务时能将各种工具集中起来使问题最小化。例如，在图 8.1 中，通过比对构建系统树

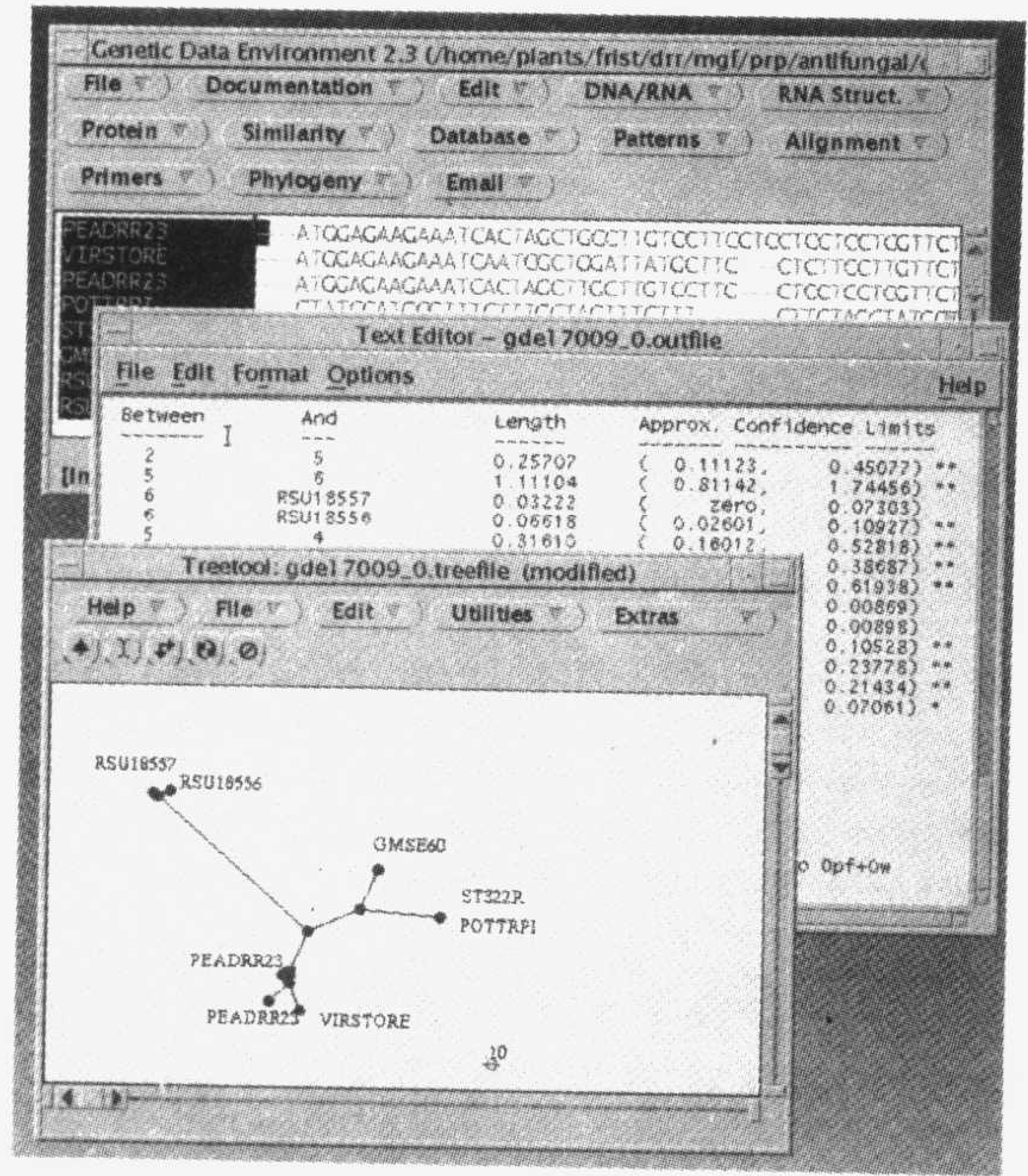


图 8.1　自动的系统树

在 GDE[2]中，选定配对的序列，调用 fastDNAml[3]产生系统树。GDE 自动调用文本编辑器和 treetool[4]来显示结果

的几个步骤就能够自动完成。对于BIRCH中目前运行的300+程序，我们没有对它的安装提供详细的说明，但对建立、维持一个不断发展的序列分析系统的各种策略和技巧作了简要的说明。

8.2 硬件、软件和技术

8.2.1 硬件和操作系统

目前BIRCH能在安装了Solaris 2.5的Sun工作站点上执行。由于它使用的源代码能应用于大多数免费软件程序上，因而可以在其他的平台上进行再编译。在BIRCH中，几乎所有的程序都能在LINUX下运行，也有许多在其他的UNIX平台上运行，若想在已建好的多用户系统上建立BIRCH，则需要某一终端或PC界面的模拟终端。当以你的账户登录时，所有的程序都将在服务器上运行，并在某终端视窗显示。终端类似于一台PC机，且每台机器有不同的软件、数据和硬件，你能从任何终端登录到服务器上。终端视窗能执行网络计算，更详细的描述见参考文献[5]。

8.2.2 软件

要求有AC编译程序，最好是GNU C以及网页浏览器，建议使用Netscape Communicator，因为它带有可视的HTML编辑器。

8.2.3 技术

要求具备UNIX[6]相关知识，有编程的经验和编写HTML语言的能力，若你对这些都不了解，那现在就是获得这些技能的好机会。

8.3 安装BIRCH

8.3.1 创建一个管理员账户

不要在自己的私人账户上安装BIRCH，而要为它创建一个独立的账户，这样所有BIRCH的目录就能形成一个独立单元。在你的私人账户上，你是作为另外的用户，这也是检验任一程序能否在任何用户上运行的唯一方法。作为管理员，应该避免在私人账户上进行初步安装。另外即使你具有使用根目录的权利，也应避免在根目录下工作，但系统工作除外(如非BIRCH)。

8.3.2 创建子目录分级结构

创建BIRCH目录，目前在我们的站点上，它们的大小如下：

GenBank：GenBank DNA 核酸数据库(5.9G，106 版本，1998 年 4 月)。

PIR：183M，55 版本，1998 年 2 月。

admin：管理文件和脚本(0.4M)。

bin：可执行文件(60M)。

dat：程序的特殊数据文件(如 scoring matricies)(5M)。

doc：文本文件(18M)。

install：安装软件的工作目录。

java：Java 类程。

manl：UNIX 手册页格式文本。

ncbi：国家生物技术信息中心(NCBI)服务器程序目录(1.7M)。

public_html：BIRCH Web 站点目录(2.5M)。

这些目录分级结构可以从 http://home.cc.umanitoba.ca/~psgendb/build/build.html 下载。所有这些目录均为管理 BIRCH 的一系列外壳脚本和数据文件。为了确保安装的软件为最新版本，最好是在建立本地的 BIRCH 站点时，下载每一个程序和软件包。

这些目录都应当安装在 BIRCH 的管理目录$HOME 下。在我们的系统上，$HOME 目录为/home/psgendb。我们设置了环境变量$db 来保存此路径。当执行某一命令时，外壳将以/home/psgendb 代替$db。因此，系统上的任一用户可以键入$db/admin 来代替键入管理目录/home/psgendb/admin。

无论任何时候你安装程序软件包时，都应在 doc 和 dat 目录下为软件包、文件和数据建立子目录。事实上，当你开始任何一项新计划时，都应建立一个特殊的目录，哪怕只是临时的，这是一个很好的工作习惯。

对于这些目录，以下有几点重要的规则：

(1) 所有的目录和文件，包括 BIRCH 的$HOME 目录，必须是全球皆可读。

(2) 所有的目录，必须是全球皆可执行。

(3) 所有的程序，必须是全球皆可执行。

8.3.3 配置管理文件

在我们的站点上，BIRCH 包含了 300 多个程序和 2 个主要的数据库。这些程序是由不同的著者在不同的平台用不同的文本格式和语言编写的。在多数情况下，他们需要知道数据文件、数据库或运行时间(runtime)库的位点。若要为每个用户安装这一切，而且每安装或上载一个新程序用户就得改变这一切，那样，所有的工作将无法进行。幸运的是，对于这些问题已有解决的方法，即所有的配置从$db/admin 读取，注意绝不能违背这个规则！用户无须对自己的账户配置任何东西。(在 Manitoba 大学，BIRCH 有 140 多个用户，设想必须为每个用户改变配置的情况！)

在 BIRCH 中，有 2 个文件含有程序需要的配置。用户一登录，$db/admin/login. source 包含的命令就被执行。最重要的命令是将$db/bin 添加到用户的$PATH 环境变量中。当外壳读取命令时，第一个非空格字符串被解释为命令的名字。外壳能搜索列于$PATH 上每一个目录中的可执行文件，这样若所有的 BIRCH 程序都在$db/bin 中，我们就必须将$db/bin 加入到$PATH，用户才能运行任一程序。另外，login. source 具有一个打印$db/admin/Login_Message 中内容的命令，一个向 BIRCH 用户发表简短告示的文件。

每当一新的外壳开始时，比如打开一个窗口或运行一个程序，就必须执行$db/admin/cshrc. source 中所包含的命令。事实上，所有的程序配置都已经限定。多数包含命令的文件需要设置环境变量，NCBI 的配置就如下：

```
# Environment variables for sequence work.
# Upper and lowercase are supported.
setenv  DB         /home/psgendb
setenv  db         $DB
setenv  DATA       $DB/dat
setenv  data       $DATA
setenv  DAT        $DATA
setenv  dat        $DATA
setenv  GENBANK    $DB/GenBank
setenv  gb         $GENBANK
```

这里，我们限定了$db，然后利用它去建立其他的环境变量来说明数据文件保存的地方。

每个程序或软件包也都有特殊的配置，如：

```
# NCBI
setenv   NCBI      $db/ncbi
alias    entrez    Nentrez
```

setenv 指示 NCBI 程序在哪里寻找必需的目录，当用户键入“entrez”时，别名(alias)行指示外壳应当运行 entrez 的网络版本(Nentrez)。

要使用 BIRCH，用户必须运行$db/admin/newuser。script 会将一行文字

```
source/home/psgendb/admin/login.source
```

添加至用户的.login 文件，同时将

```
source /home/psgendb/admin/cshrc.source
```

添加至.cshrc 文件。当用户登录或开始一新的外壳时，这两行就能使得.source 中所有的命令都能被分别执行。用这种方法，$db/admin 目录下的.source 文件的任何改变或添加都能对每一用户自动产生作用，而 BIRCH 管理员无须对用户的账户作任何变动。实际上，要想顺利运行的话，login. source 和 cshrc. source 必须

作些修改而得以反映本地目录结构，同时安装软件。例如，在你的本地 BIRCH $HOME 目录中，必须改变$db 环境变量。在 login. source 和 cshrc. source 中，涉及一些尚未安装的程序或数据库，本应该加以注解，而 BIRCH 是不断发展的，这些就无法注解了。

8.3.4 创建网站站点

将 BIRCH Web 站点视为你正要建立的概念模型，在这里，它告诉了用户怎样使用这一系统，以及哪些是有用的，但 BIRCH 的复杂性要求建立一个结构清晰的路线图。若你的屏幕上不断运行 Netscape 的复本，你就能创建 Web 页并修改它们。由于在 Manitoba 大学的站点[1]上 BIRCH 早已存在，所以你可以走捷径，通过下载网页并按你的需要修改。在后面章节里的描述，都假设存在一个网页，名称为 programs. html(见 `http://home.cc.umanitoba.ca/~psgendb/programs.html`)，此页包含了所有文档的链接，且文档都已分门别类。

8.4 建立 BIRCH

此节描述的是安装几个软件包的全过程，每个软件挑选了一些细致的问题来加以说明，这些问题与利用程序为分散用户工作有关。此节的目的就是为这套应用相当广泛的程序快速工作起来提供一条捷径，程序的核心是为满足你的本地用户的需要而建立合适的系统服务的。简单来说，下载程序的服务器地址在后面的文献中皆有，且程序通常带有比这里更详细的安装说明书。

就程序的安装而言，最好是在管理员账号下安装，而在私人账号上检验，这样在账号中来往就方便一些。对于这个问题有两种方法，理想的途径需要有联在一起的两台终端，每台用不同的账号登录，这通常难以实现；另一途径是在你的私人账号上运行一个窗口对话，但在一个或多个窗口中以管理员账户登录。比如开一命令窗口，然后用 telnet 登录你的管理员账号，对于简单的任务，以管理员账号登录，保持一个或多个 telnet 对话，每一个工作目录有一个即可，若想在管理员账号上运行 X11 程序而在你的私人账号上显示出脚本$db/admin/xdisplay，就需要对你自已的站点作调整，如果你使用的是 CDE 台式管理器，来自你的私人账户上的所有窗口列于一个屏幕，而来自管理员账户上的所有窗口列于另一独立的屏幕，这样就不至于混乱，同时，BIRCH 的 `newuser` 脚本能促使你的 UNIX 提示服务器名称和当前目录，这样就有利于弄清哪个窗口属于哪个账户。

8.4.1 安装 readseq[7]

序列软件最大的一个问题就是所用的文件格式太多，readseq 是一个将一种格式(如 GenBank)转换成另一种格式(如 GCG)的程序，它的源代码和文档能以一种外

壳档案文件(readseq. shar)形式下载。若想在档案文件中重新创建文件，可键入 sh readseq.shar。若想你的平台能编译 readseq，可通过键入 make 来创建可执行文件 readseq，从而使这个文件全球可读。

```
chmod a+rx readseq
```

然后将其移至 bin 目录

```
mv readseq & db/bin
```

同时为帮助文件创建一个目录

```
mkdir $doc/readseq
chmod a+rx $doc/readseq
```

再将帮助文件移至此目录

```
mv readseq.help $doc/readseq/readseq.asc
```

通常对于其他的软件包都不重命名文件名，然而，网页浏览器通常随处理的文件扩展名不同而变化，因而对所有的 ASCII 文件最好统一文件扩展名，这里采用“.asc”，最后将 readseq. asc 链接到 programs. html 上。

在私人账号上读取 readseq 文件，并检验各程序，例如，若有一个名为 PEADRRA.gen 的 GenBank 文件，键入

readseq-p-oPEADRRA.wrp-fPearson<PEADRRA.gen 时，将创立一个 Pearson/FASTA 格式的文件，文件名为 PEADRRA. wrp。read seq 最初是由 VMS 发展而来的，因此对使用 UNIX 输入重定向符号“<”的程序来说，管道输入必须利用“-p”开关。

8.4.2 安装 FSAP[8, 9]

许多程序都以软件包形式存在。FSAP 软件包包含许多常见的序列任务程序(如打印序列、翻译、限制性位点搜索)，它们通过相互作用的文本界面菜单方式运行。这样，软件包就以扩展名为.tar 的档案文件下载，fsap. tar. Z。若要建 FSAP 的目录分级结构，首先应将文件解压缩。

```
uncompress fsap.tar.Z
```

然后键入 tar xvf fsap.tar

从而创建 fsap 目录，若你在 fsap 目录下键入 ls -1，将见到如下文字：

```
drwx——2 frist drr       512 Jun 4 1996 GDE/
-rw—— 1 first drr      7041 May 3 1996 INSTALL.doc
-rw—— 1 frist drr       970 May 3 1996 RELEASE.NOTES
drwx——2 frist drr       512 May 3 1996 bin/
drwx——2 frist drr       512 May 3 1996 dat/
drwx——2 frist drr       512 Mar 6 18:56 doc/
drwx——4 frist drr       512 May 3 1996 src/
```

```
drwx——2 frist drr      512 May 3 1996 src.c
drwx——2 frist drr     1024 May 3 1996 test/
```

INSTALL. doc 文件带有逐步安装的说明书，src.c 具有 C 语言源代码，产生可执行码，doc 和 dat 分别含有程序使用的文档和数据文件。

Genetic Data Environment(GDE)具有菜单目录和 c-外壳脚本，从而使这些程序能通过 GDE 运行。test 是运行某脚本中的目录，它检验所有的程序并确保它们在你的系统上发挥作用，许多软件包都有 test 脚本。当你成功地检验各程序后，安装起来就容易了，将 fsap/bin 的内容复制到$db/bin 上，fsap/dat 至$dat/fsap/dat 上，doc 至$doc/fsap/doc 上，并确保这些文本文件都链接到 programs. html 上。

之后，登录你的私人账号，并测试这些程序，首先要测试的是 numseq，与$doc/fsap/numseq. asc 中描述的一样，任何 GenBank 中的文件都能满足测试这个程序的要求。

8.4.3 安装 FASTA[10]

FASTA 软件包提供的程序具有碱基对和数据库序列比较的功能，编辑由 UNIX 中的“make”命令完成，安装很简单，复制可执行文件至$db/bin 即可，它是最易安装的软件包之一。可是此文件为 UNIX 手册页格式。BIRCH 建有手册页目录，为$db/manl。在此目录下的所有文件都以“name. l”的形式存在(这里“l”代表本地)，在 login. source，此行 setenv MANPATH $MANPATH\:$DB 指示 UNIX 在此目录下和其他目录，特别是系统目录$MANPATH 下找寻手册页。

例如，想读取名为 align 的文件，用户键入“man align”，文件$db/manl/align. l 就列出来。

此外它对从手册页中创建 ASCII 文件也有用，并能通过网页浏览器展现出来。将 align.l 转换成 ASCII 文件，键入 man align>$ doc/fasta/align.asc 即可，记住使此文件为全球可读并在 programs. html 中建立链接。

8.4.4 安装 GDE[2]

GDE 是一个运行其他程序的程序，如图 8.1 所示，GDE 将多序列编辑器和下拉菜单结合起来。例如，它的 Similarity 菜单包含的内容就与 FASTA 的相似，GDE 独特的地方就是添加菜单项时不需要再编译。当 GDE 启动时，读名为$GDE_HELP_DIR/.GDEmenus 的文件，指定每个菜单的内容，同时执行各个命令运行每个程序，如图 8.2 显示的 lfasta 菜单。

在.GDEmenus，创建 GDE 菜单方式如下：

```
# ———————— LFASTA(7/26/95) ————————————
item:LFASTA - Fast local alignment
itemmethod:(sed "S/[#%]/>/" <in1>in1.tmp;readseq
```

```
in1.tmp -i1 -f8 > in1.seq1;readseq in1.tmp -i2
-f8 > in1.seq2;lfasta -w $RESPERLINE $MARKX -d
$NUMOFALN $MATRIX in1.seql in1.seq2 $KTUP >
in1.out;
fastaout.csh $MARKX in1.out;rm in1*)&
itemhelp:FASTA/fasta.asc
```

最重要的行是 itemmethod，它包含一串命令需要运行，例如，readseq 将挑选的序列转换成 FASTA 格式，命令中插入了 arguments，前面有一符号“$”，lfasta 中都有几行详细说明，如图 8.2 中显示的下拉菜单 DISPLAY。

```
arg:MARKX
arglabel:DISPLAY
argtype:choice_menu
argchoice:Identity="colon"Cons.repl.="."Mismatch=" ":-m 0
argchoice:Identity=" "Cons.repl.="x"Mismatch="X":-m 1
argchoice:Print only 1st seq;Identity="."Mismatch="residue":
-m 2
argchoice:graph of conserved positions:-m 4
argvalue:0
```

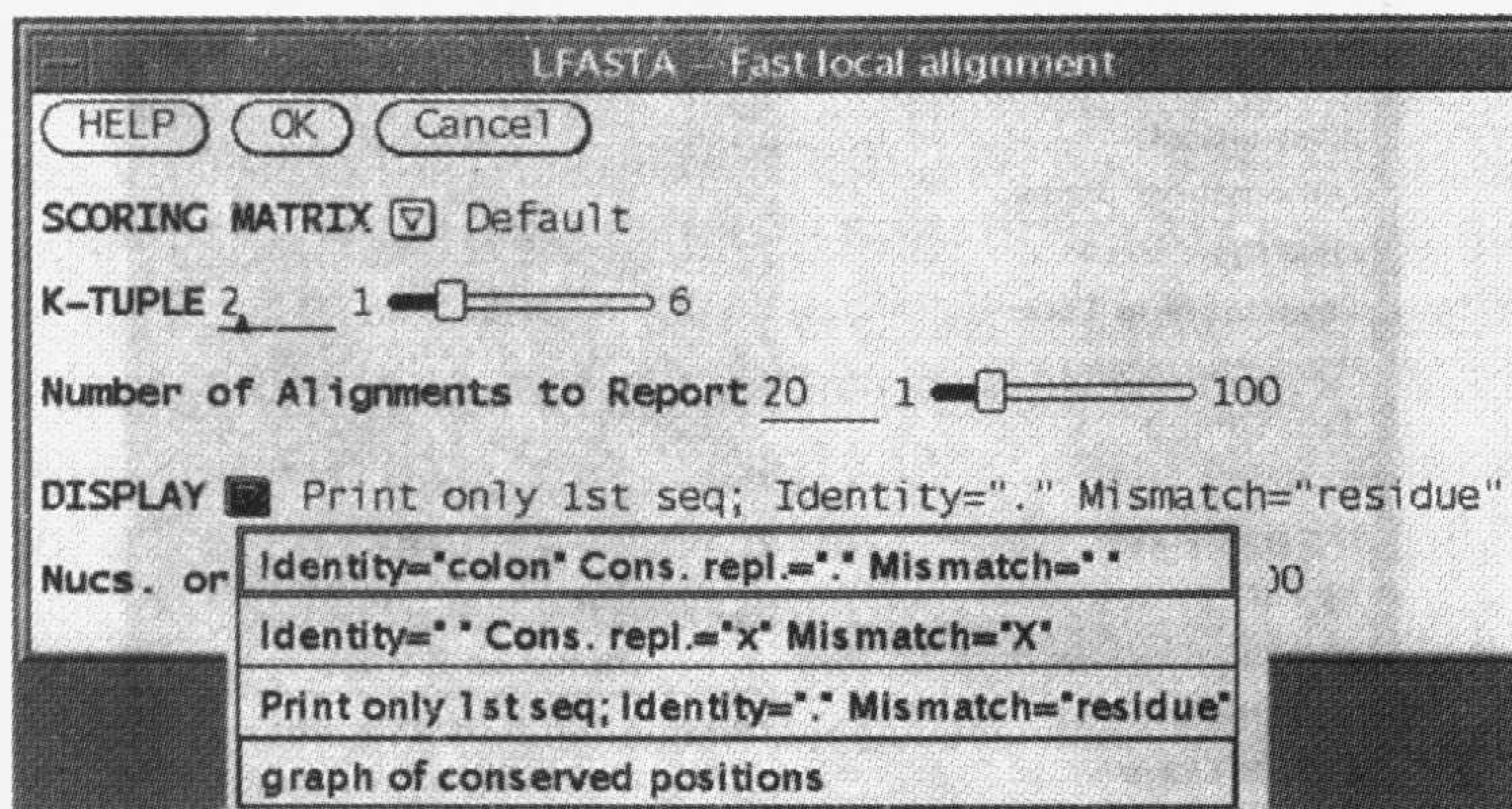


图 8.2 GDE 菜单实例

显示了下拉菜单和选项卡，当 lfasta 从 Similarity 菜单被选取时这个菜单就显示出来了

对于那些从命令行中接收所有信息的程序，如 readseq 程序，GDE 运行它们是最容易的，即使是那些相关联的还需要用输入的程序也能由 GDE 运行，例如，要运行 numseq，GDE 将由菜单设置的参数输送至名为 numseq. csh 的文件，此文件读取参数后，并对用户键入的指令在 numseq 提示下产生响应。由于新程序容易添加至 GDE 菜单中，使得 GDE 成为 BIRCH 得以运行的基础。BIRCH 中的 A. GDEmenus 文件和大量程序运行所必需的附属外壳脚本能从参考文献[1]中下载。

8.4.5 安装 NENTREZ[11]、SEQUIN[12]、BLASTCLI[13]和 Cn3D[14]

NCBI 由网络代理/服务器组成。Nentrez 是运行在你的电脑上的一个代理，它可以从 NCBI 服务器上进行文件搜索和序列恢复。它的辅助应用程序 Cn3D 能下载和显示来自 NCBI 中蛋白质数据库中的三维结构。sequin 自动执行评论新序列并将其送入 GenBank 中，也能从 NCBI 中下载序列。blastcli 是本地代理，将序列递呈至 NCBI BLAST 服务器中。

由于这些程序共享一个含配置文件和数据文件的目录，因而安装很快。通常，只需复制可执行文件至$db/bin 上并运行网络代理配置程序 netentcf 即可，这些程序的所有文件和目录都应当在$NCBI 指定的目录$db/admin. cshrc. source 下。当你首次运行 Nentrez 时，将创建一个携有配置信息的文件$HOME/.ncbirc，当你将此文件移至$NCBI 时，它就能为所有的用户所利用。

Workspace 菜单(图 8.3)能使用户非常容易地明白系统上哪些程序是可以使用的。所有的 UNIX 视窗管理都有一个可配置的 Workspace 菜单。CDE 管理器在大多 UNIX 平台上都能获得，但现在许多平台已经没有了。因此，这里为所有的 BIRCH 用户创建了一个$db/.dt/dtwmrc 文件来配置 CDE Workspace 菜单。程序被组织成

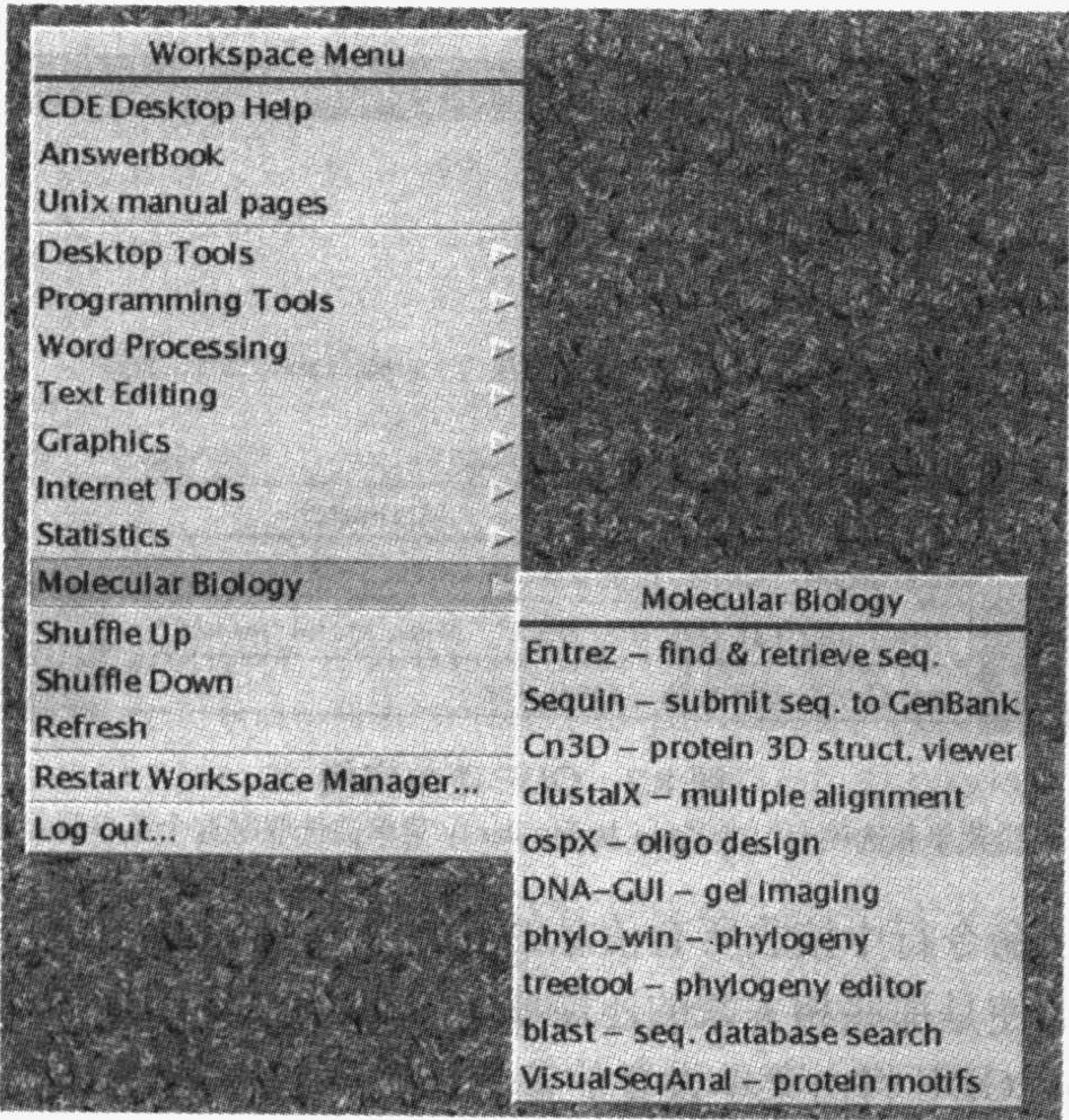


图 8.3 定制的 CDE 工作间菜单

分类好的子菜单使用户查找和启动程序更加方便

亚菜单的形式(如字处理、统计、分子生物学)。要将 sequin 添至菜单中，则"Sequin-submit seq.to GenBank"f.exec/home/psgendb/bin/sequin 必须在 dtwmrc 下，$db/bin/menusetup 这个设置脚本替代用户的 dtwmrc 文件，此文件是用一字符连接至$db/.dt/dtwmrc 的。因此，当 BIRCH 管理员更新了这个文件，所有用户得到的就是新菜单。

在此菜单中没有 GDE，这是故意遗漏的，若它在 Workspace 菜单中启动，将不能在$HOME 目录中读写文件。

在使用 GDE 时，最好为每个项目都建立一个单独的目录，例如，多基因家族的系统分析。若你切换到另一目录中，并从命令行中启动 GDE，那么所有输入和输出的文件(包括临时文件和目录)都将在此目录中出现。

8.4.6 安装 XYLEM[15]

XYLEM 是一套本地数据库管理工具，虽然当初它是为研究像系统发生这样的项目而创建 GenBank 或 PIR 文件子集而设计的，但它也能用来从这些数据库中进行关键字搜索和检索条目，features 能自动执行从大量条目中提取 GenBank 特性(如外显子、内含子、mRNA、CDS)。

XYLEM 的安装类似于 FSAP，后面章节的讲述都假定你的系统已经安装了 XYLEM。

8.4.7 安装 GenBank 和 PIR

安装像 Nentrez 和 blastcli 这样的程序，并不一定需要有序列数据库的本地版本。由于 GenBank 需要 6G 空间，因此应当深思熟虑。然而，一个数据库的本地复本还是有用的，以下几个原因可以说明：首先，网络 BLAST 程序限制了你对 GenBank 中特殊部门的数据库搜索，而本地数据库搜索程序，如 FASTA 能满足你的特殊需要。最后，Nentrez 不允许对修改了的序列输入一组新的目录号。

通过外壳脚本 gbupdate 和 pirupdate 可以完成 GenBank 和 PIR 的自动下载和格式化。例如，要下载的 GenBank 的文件存在于$gb/master. filelist 中，那么在高峰期后开始下载，就可以使用“at”命令：

```
at 7pm
at>nice gbupdate master.filelist
at><ctrl>-D
```

UNIX 的 nice 命令在执行此功能时有一点优先权，因此，在实际工作时(如在屏幕上移动窗口)速度并不减慢。gbupdate 能下载每个文件，并确认下载的文件与原来的未解压缩的文件大小一样，而对于 GenBank 中的序列文件，可以运行 splitdb[15]将它分解成注解、序列和索引。序列文件为 FASTA 格式，序列和注解被分成独立的文件其速度取决于 FASTA 搜索序列和 findkey 搜索注解的速度[15]，

fetch 能恢复 GenBank 条目，重新连接注解和序列以便重建原来的条目。

GenBank 比较大，现在被分解成了几个文件。EST 就被分解成 22 个文件，106 版本是 gbest1~gbest22，反应这些变化的文件 master. file list 每次下载都需要更新。而 fetch 和 findkey 能自动检测到各类文件被分解。

8.5 培训用户

对于 BIRCH 来说，人们通常不能读懂它，而保持它的完整性和连续性却非常关键。亲自培训章节对于群体用户来说可能更有价值，它既教会人们怎样使用 BIRCH，也培养了用户之间的团结互助。

由于大多数 BIRCH 用户不了解 UNIX，因而想在这两方面都学好似乎没有希望。但是，在每年我讲授的实验室课程细胞遗传学[16]中，一些没有 UNIX 经验或不了解生物信息学背景知识的学生在学过两次以上的手动培训课后，就能完成一个简单的序列设计。在第二节结束后，给每一个学生一段 300~400bp 来自 GenBank 的编码蛋白的未知序列，要求他们必须能用 FASTA 鉴别、检索它的原序列，鉴别未知序列的编码序列并打印整个编码序列，用正确的读框翻译。

此节的进程如下：

(1) 快速示范(30min)：开始先给学生一个示范，使他们对整个事情是怎么一回事有一点概念。使用的 X-终端与 1024×768 投影仪相连。我简单的解释了 X-Windows 是怎样工作的和 CDE 桌面的基础，并分别用行命令程序和通过 GDE 运行的程序向他们示范了怎样分析序列。

(2) UNIX、CDE 和简单的序列任务的亲自示范(2h)：示范需要一步一步进行，以确保在课程往下进行以前都能成功地完成每一步。当他们遇到困难时给予帮助是很重要的。首先，学生运行$db/admin 中的安装脚本目录 newuser 和 menusetup，然后介绍 CDE 的基础以及使用网页浏览器读文件。接下来，学习从命令行中运行 numseq 而分析序列。numseq 能说明 DAN 序列单链或双链的扩增，链状和环状分子的区别以及怎样翻译序列，然后发射 GDE 并重复某些同样的工作，从 GDE 菜单中运行 numseq。(虽然，图形界面较好，但最好教会一些命令行的技能，这样做的话会对计算机实际所做的事有深入的了解。)这节以对 GenBank 的快速讨论结束，学生学会用键盘搜索序列和收回它们。

(3) 相似性搜索和数据库(2h)：此节开始对点矩阵[9]和整体[10, 17]相似性搜索的原理进行了简短的公开的讨论。强调了 look-up 表和优化调试机器设备的概念。并学习了使用 GDE 对几组相关序列进行碱基对比较，d4hom[9]进行点矩阵搜索和 align[10]进行调试机器整体设备。最后，学习运用 fasta 在 GenBank 中搜索 DNA 序列。

(廖晓萍　译)

参考文献

[1] Fristensky, B. BIRCH.http://home.cc.umanitoba.ca/~psgendb.

[2] Smith, S., Overbeek, R., Woese, C. R., Gilbert, W., and Gillevet, P. M. (1994) The tenetic data environment: an expandable GUI for multiple sequence analysis. *Comp. Appl. Biosci.* (now *Bioinormatics*) **10,** 671-675. ftp://megasun.bch.umontreal.ca/pub/gde/.

[3] Olsen, G. J., Matsuda, H., Hagstrom, R., and Overbeek, R. (1994) FastDNAml: a tool for construction of phylogenetic trees of DNA sequences using maximum likelihood. *Comput. Applic. Biosci.* (now *Bioinformatics*) **10,** 41-48.

[4] Maciukenas, M. (1994) Treetool. ftp://rdp.life.uiuc.edu/rdp/programs/TreeTool/.

[5] Fristensky, B. (1999) Network computing: restructuring how scientists use computers, and what we get out of them, in *Methods in Molecular Biology* vol. 132.*Bioinformatics Methods and Protocols* (Misener, S. and Krawetz, S. eds.), Chapter 22. Humana Press, Totowa, NJ.

[6] Sobell, M. G. (1995) A *Practical Guide to the UNIX System.* Addison-Wesley Publishers.

[7] Gilbert, D. (1993) http://iubio.bio.indiana.edu/soft/molbio/readseq/.

[8] Fristensky, B., Lis, J. T., and Wu, R. (1982) Portable microcomputer software for nucleotide sequence analysis. *Nucl. Acids Res.* 10, 6451-6463. http://home.cc.umanitoba.ca/~psgendb/FSAP.html.

[9] Fristensky, B. (1986) Improving the efficiency of dot-matrix similarity searches through use of an oligomer table. *Nucl. Acids Res.* **14,** 597-610.

[10] Pearson, W. R. (1990) Rapid and sensitive sequence comparison with FASTP and FASTA. *Meth. Enzymol.* **183,** 63-98. ftp://ftp.virginia.edu/pub/fasta/.

[11] NCBI. *Nentrez.* http://www.ncbi.nlm.nih.gov/Entrez/Network/nentrez.overview.html.

[12] NCBI. *Sequin.* http://www.ncbi.nlm.nih.gov/Sequin/index.html.

[13] NCBI. *Blast client.* ftp://ncbi.nlm.nih.gov/blast/network/.

[14] NCBI *Cn3D.* http://www.ncbi.nlm.nih.gov/Structure/cn3d.html.

[15] Fristensky, B. (1993) Feature expressions: creating and manipulating sequence datasets. *Nucl. Acids Res.* **21,** 5997-6003. http://home.cc.umanitoba.ca/~psgendb/XYLEM.html.

[16] Fristensky, B. *Introductory Cytogenetics,* University of Manitoba. http://www.umanitoba.ca/afs/plant_science/COURSES/CYTO/.

[17] Needleman, S. and Wunsch, C. D. (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* **48,** 443-453.

第二部分　分子生物学软件

9 Macintosh和 MS Windows计算机分子生物学方面的免费软件

Don Gibert

9.1 引言

通过网络和其他来源，你可以获得很多关于分子生物学和化学的应用程序。这些程序大部分是作为大学、国家的基金项目或非基金项目由生物学家、化学家和软件开发者编写出来的。其中的一些软件提供了对你具有简单功能但又很实用的程序，虽然可能没有经费资助。

对于生物学家来说，什么是免费？且尤为重要的是如何能获得免费的软件？如何进行安装和使用？怎样获得性能良好的程序？在本文中这样的问题都能找到答案，但本文更主要关注于如何使用 Macintosh 和 Microsoft(Wintel)操作系统的分子生物学软件。

阅读本文后，你将能通过网络浏览器上网获得许多目前运行于因特网服务器的免费程序。它们是以源程序(不能直接运行)或 UNIX 的形式的软件，或化学、医学、种群生物学、生态学等方面的软件。

9.1.1 免费或商业软件

免费的数据分析软件在科学领域是很普遍的，随着科学家们对于新算法分析要求的提高，逐步明确新的计算方法并将其应用于软件。目前，绝大多数的生物学序列分析程序是由科学家总结出来并开发成软件的。包括 FASTA、BLAST、Clustal、MFOLD、PHYLIP、Paup、CAP 等，在此不一一列出。这些程序通常状态下是免费共享的，但其缺点是使用较繁琐且不能和其他功能很好的兼容。商业软件开发商已经开发出捆绑了这些产品并增加了新的使用功能的计算软件，从而可以使消费者能够轻松的使用。

除了兼容性和用户界面外，软件开发公司还为他们的软件新增了实用的文档、电话和其他支持。如果你的经费允许，而恰巧市场上又有合适的软件包，购买软件要比下载免费程序划算得多。考虑到商业市场的现状，以及投资到广告、技术支持、软件开发等方面的开支，软件开发商的报价还是可以接受的。

对于那些缺少资金的教师、学生、科研工作者来说，下载免费的软件还是很好的选择，而且你也可以找到那些开发商销售的软件包中所没有的程序。免费软

件的另一个有利因素是，它通常包括能修改和扩充分析程序的源程序。

今天，随着生物信息学和生物计算机的飞速发展，越来越多的程序员不断地开发新的软件，其中大部分可以免费下载。你可以找到许多某些软件入门的小窍门和升级用户界面的软件。免费软件的开发仍很困难——政府一般不会为个人或某个程序员提供资金进行开发，但这些人是免费软件的主要提供者。共享自己付费的软件还没有在广大的用户中间形成共识。而对于科学领域来说潜在的市场太小，这就是为什么我们所看到的昂贵的商业软件包和下载的免费软件有巨大差异的原因。

免费程序可用性是可变的，但它有赖于你的计算机和你的需要。需要注意的是：较陈旧的程序往往不能在新型计算机上运行。对程序漏洞和局限的容忍是必要的，并且要对学习班它们如何工作要有自信心。通常联系不到作者或作者没时间向别人提供他们所共享的产品相关的信息。

9.1.2 网络资源

如今，因特网已被普遍使用和大为推广，特别是在科学领域中尤为如此，这使得你已经获得或将获得这项优势。这在几年前还几乎是不可能的事，但自从1980年以来，免费的生物软件已经在因特网上大为普及，因为它不需要向软件制作者付费。原来只有少数几个在网上共享的生物学软件，现在，制作人更愿意推广他们的软件而不是网络服务器，因为这样较为简便。收集这些软件的档案室仍在提供这些收集品方面起着重要作用，这些一直保存的文档在使用时对人们非常有用。

两处常用的分子生物学档案馆是印第安纳州大学的 IUBio Archive 和欧洲生物信息学会(European Bioinformatics Institute，EBI)，以及他们的网站因特网资源定位(Internet Resource Locators，URLs)：

EBI 在 http://www.ebi.ac.uk/或者 ftp://ftp.ebi.ac.uk/。

EBI 是 EMBI 数据库和其他数据库的主页，也是大型分子生物学的网站，包括分子生物学软件的下载网站，其中包括常用的 Bio Catalog of software(http://www.ebi.ac.uk/biocat/biocat.html)。

IUBio 在 http 或 ftp://iubio.bio.indiana.edu/。

IUBio Archive 文档是一个大型的生物软件收藏主页，它为当前 GenBank、SwissProt 和 PIR 数据库的关键词检索提供服务，它还提供实用的 Bionet 网络信息文档。从 1989 年开始它就作为一个用户和制造商支持单个用户的文档运行至今。在 IUBio 中的分子生物学软件收藏也被镜像到世界各地，包括芬兰、瑞士、日本、英国、法国、西班牙和爱尔兰。下面有这些软件的 URLs 的列表。

许多著名的其他的因特网服务器包括：

http://www.ncbi.nlm.nih.gov/—— Entrez 软件(数据库查寻)的最新

访问的主页地址，也是为在 GenBank、Macaw(多排列)和其他发布个人的序列的 Sequin 的主页地址。

http://kiwi.imgen.bcm.tmc.edu:8088/search-launcher/launcher. html——Baylor 医科大学的 Search Launcher 序列分析方法选择的主页。要推荐的是他们的 Search Launcher 程序(在 Macintosh 和 Wintel 系统中可使用的一种 Perl 程序)是用来收集多线程访问这个有价值的资源的。

http://expasy.hcuge.ch/——SwissProt、PROSITE 的主页，有用的数据库分析功能(尤其是对于蛋白质数据)集合的主页。

9.1.3 获得软件

文件传输协议(ftp)是专门给文件传输指定的因特网服务。它是普遍采用的超文件传输协议(http)或网页、因特网方法的前身。概括地说，ftp 对于传输大型文件如软件包仍是优于网页浏览的方式。在 MacOS 中流行和简单的使用 FTP 程序是 Dartmouth 大学 Jim Matthews 编写的 Fetch 程序。你可以在 http://www.dartmouth.edu/netsoftware/fetch.html 中找到更多的信息。常用的 Netscape 或其他网页浏览器也会使你通过 http 或 ftp 方式获得软件。

基本上，软件是从文档中以加密的格式被储存和传输给你。当前大部分获得的软件，如 Fetch 和 Netscape 将为你自动解密。尽管如此，其他软件也用于解密或需要解密。大部分因特网软件包和服务器会使你在需要的地方运行。如果不确定，你的当地图书馆或书店将会有一些关于如何使用因特网服务和基本软件文档的书。

当今获得软件的主要技巧是知道在什么地方查找。除了这篇文章，你可以使用别人制作的网页连接页，可能有比本文作者提到的更好的前景。同时也应注意通用的搜索服务，如 Yahoo(http://www.yahoo.com)、AltaVista(http://www.altavista.digital.com)、Lycos(http://www. lycos. com)或其他。

在头脑中要有这样的意识：软件是适时的。如果你所试用的版本在某些情况下不能使用时，应立刻或很快得到一个较新的版本。本作者偏爱的或主页服务器是检验更新的版本的最好场所，因为档案服务器通常没有当前的版本。

9.1.4 安装和使用

以前，在你的计算机上安装免费软件的步骤是由易到难。通常是包含说明的，但总是不够详细，或不能覆盖我们所能碰到的所有问题。安装软件通常是商业软件，而非免费软件。

向编写程序的作者报告你的问题和改进软件的建议。然而，频繁地向程序制作人求助安装并不总是好的解决方法，所以注意他们的写作前的构思对安装向导是必要的。

专门的软件，尤其是那些用 Java 或 Perl 编写的，还需要你获得并安装别的免费的软件。对于 Java，这变得不成问题，因为新的 MacOS 和 UNIX 系统正使之成为系统的一部分。

9.1.5 多平台软件

在 1998 年以前，在 IUBio 和其他生物网页服务器上因特网浏览器协议的数目中，30%~50%的生物学家使用 Mac 计算机，40%~70%的生物学家使用 Wintel 系统，同时大约小于 10%的生物学家使用 X-Windows 系统作为他们的工作站(尽管许多人把 UNIX 或 VMS 用于其他的事情)。生物学为软件对计算系统的选择和需要保留多个种类。许多人根据软件在不同系统中的运行使用或拥有多个操作系统，而且，有允许在 Macintosh 运行大多数 MS Windows 和 MS DOS 软件，以及在 Wintel 系统中运行一些 Macintosh 程序的商业模仿程序。

一些软件可以在多个操作系统中运行。这可能是一个某些人梦寐以求的圣杯，这些人开发科学应用程序——人们希望程序能被任何需要它的人在任何计算机系统上使用，但是编写多平台软件并不容易。除最近的 Java 的成功外，还没有容易且好的方法可以并不昂贵的代价做到这样。甚至大的商业开发者尽了很大的努力也并不一定总会获得成功。

随着我们所期待的图像界面软件的出现，多平台程序可能在你的特定的计算机系统中看起来并不是十分正确，即使它们按规定的操作进行。在更为普通的没有图像界面的软件的情况下，你可能需要把额外的时间用于学习它的命令行或菜单驱动的语法。

在 Sun Microsystems(http://www.javasoft.com)上诞生的新的 Java 开发系统正提供开发在普通系统中能很好运行的有用的软件的方法。当前 Java 软件常比用 C++、C 或其他语言写的软件要慢。我们可期望在未来看到更多的用 Java 写的用于生物计算的软件。来自 Licor(www.licor.com)的新的序列分析软件包就是一个例子。

9.1.6 客户-服务器生物序列分析软件

有许多开发者从事将用户界面与分析程序分离的工作，且我认为这是一种使程序更易于使用的有效方法，是客户-服务器软件设计的基础。其简单的例子，如当前为不同数据分析而大量存在的网页接口。

作者自己改进的 SeqPup 就采用了这个方法：它允许你使用你所需要的分析软件，Clustal、CAP、tacg、fastDNAml 或其他，运行在你自己的计算机或服务器计算机上。SeqPup 为编辑序列、基本操作和排列，以及高级显示和输出参数等提供图像界面和标准用户接口的方法；它也能根据你需要的安装方式连接到分析机上。这些分析程序给复杂的数据分析算法加密，但除命令行参数外基本上没有用户接

口。随着用户程序，如 SeqPup 的出现，这些程序的使用得到简化且便于你组织序列数据。

在 EMBL/EBI 的 Martin Senger 正忙于被称为 AppLab(http://industry.ebi.ac.uk/applab/)的序列分析软件的基本 CORBA 接口，这是一种相似的方法。

Peter Rice 的 EMBOSS(http://www.sanger.ac.uk/Software/EMBOSS/)将是一个可自由分配的分析程序系列，它仍处在发展之中。它将用命令行接口运行于 UNIX 服务器计算机。EMBOSS 将会包括不同的序列分析主题，也将包括许多的努力以供其他公共领域的简单合并。分析包括：搜索序列格式的快速数据库检索、序列重叠、简单的和特种重复的鉴别、核苷酸序列分析、小基因组的密码子使用的分析、基因组测序的基因鉴别工具、在大规模序列组中的序列类型快速鉴别，以及蛋白质基序鉴别和发行的工具。

这些是基本的序列分析机，一些有较好用户接口的客户端程序，如 SeqPup、Java applets 或网页形式，可能是你用以进行分析的程序。

9.2 免费软件的最显著部分

9.2.1 Clustal 序列比较

Clustal 提供自动多序列比对。其当前的版本叫 Clustal W，它可用在 MacOS、Wintel、UNIX 和 VMS 计算上。许多核苷酸或氨基酸序列的同时比对是现在分子生物学的基本工具。多比对用以找到描述蛋白质家族特征的诊断类型；用以检测和显示新序列与现存序列家族之间的同族关系；用以帮助预测新序列的二级和三级结构；用以提供 PCR 的寡核苷酸引物；用以作为分子进化分析的关键前序。Cplt 程序很好地满足了这个需要，它可在 ftp://ftp-igbmc.u-strasbg.fr/pub/-Clustal W 以及 EBI 和 IUBio 档案中得到。有一个伴随程序 Clustal X，它为 Clustal 提供图形界面，Clustal 可以从其他序列编辑器(如 SeqPup)上使用。

9.2.2 搜索基因组数据的 Entrez

Entrez 程序是用于基因序列数据和 MEDLINE 文献的关键词搜索的。它已由在国家生物技术信息中心(NCBI)的编程组写好。它从 http 或 ftp://ncbi.nlm.nih.gov/中得到。Entrez 运行于各种计算机系统。Entrez 的一个优势是它包括了 MEDLINE(它包含提交了序列数据的摘要)。NCBI 的网页服务也通过你的网页浏览器提供了 Entrez 类型的能力。在 NCBI 开发的 Entrez 程序资源已经为其他生物科学应用程序提供软件框架，包括 SeqPup 和 Clustal X。

9.2.3 用于图像分析的 NIH Image

一个非常有用的基本图像分析的 Macintosh 程序是 NIH Image，它是由 Wayne Rasband 编写的。Image 可用以测量面积、平均密度、引力中心和影响用户指定区域的位角。它也做少量的自动分析和能被用以测量路径长度和角度。测量结果能被打印，输出到文本文件，或被拷贝到剪贴板。结果可被用作真实的值标准。在 ftp://zippy.nimh.nih.gov/pub/nih-imge/，或 http://rsb.info.nih.gov/nih-image/中能可找到 Image。现有用于 MS Windows 的版本，它来自 http://www.scioncorp.com/。这个程序的使用在第 14 章有描述。

9.2.4 用于发展史分析的 PHYLIP

广泛使用的 Phylogeny Inference Package(PHYLIP)来自 Joseph Felsenstein，它是用以推测发展史(进化树)的程序包，它可以在尽可能多的不同的计算机系统上运行。它包括分析 DNA 和蛋白质序列、限制酶切位点、距离矩阵和基因频率、数量和离散的特征和进化树的标图。使用的算法包括简约、最大相似性、邻域连接和一些其他的方法。分析的精确控制的许多参数是可得到的。PHYLIP 的主页在 http 或 ftp://evolution.genetics.washington.edu/。该程序的使用在第 12 章中有讨论。

9.2.5 用于分子模拟的 RasMol

RasMol 是一个广泛使用的、免费的蛋白质、核苷酸和小分子可视化的分子图像程序。这个程序旨在显示、讲授和出版性质图像的产生。RasMol 运行于所有普通计算机系统。该程序读出分子的坐标和交互地用各种颜色图显示分子和分子表达，包括深度提示线架、空间填充球体、球形的和棒状的立体线状的生物分子带状物、原子标签和点状的表面。RasMol 的主页是在 ftp://ftp.dcs.ed.ac.uk/pub/rasmol/。

9.2.6 用于序列编辑的 SeqPup

SeqPup 和其前的 SeqApp，是生物序列编辑器和分析程序。它们包括对网络服务和额外分析程序的连接。SeqPup 在普通计算机系统都可用，因为它使用最新的 Java 语言。

特征包括多序列比对和单序列编辑、读和写多序列文件格式、比对结果以及带盒式和阴影区的序列的精美打印、序列特征编辑、操作和打印标记，一致性、反向互补、长度/同源性和 DNA 与蛋白质转换。打印文件格式包括 PICT、PostScript 和 GIF。

用户定义的连接外部分析程序，包括 Clustal W 多序列比对，还有 CAP contig

汇编，tacg 酶切图谱和 fastDNAml 进化树分析，其他的也可被加上。人们能将其运行于自己的计算机或因特网服务器计算机，使用新的 CORBA 协议(www.corba.org)。因特网序列分析服务包括获得序列，使用 SRS 关键词搜索和执行 NCBI BLAST 同源性搜索。

SeqPup 的主页是 http://iubio.bio.indiana.edu/soft/molbio/seqpup/。要注意的是这个应用程序正在不断地改进之中，它有一些小毛病。SeqApp 是仅能在 Macintosh 上使用的 SeqPup 的前版。许多人当前发现一个比 SeqPup 更有用的程序，速度更快，但缺少 SeqPup 最新特征。

9.3 软件使用问题

9.3.1 授权与大众范围

许多免费软件是由作者或资助者授权的，他们保留所有的权力。特地授予你免费使用软件的权力，可能只做非商业用途。商品化的授权软件的使用或其他的赢利性的使用需要作者的同意。许多免费软件的程序带有源程序，所以你可以修改和扩充它。这对于让高级用户做一必要的分析是有很多好处的，但记住它是不能用于商品产品的。如果作者明显把他的作品放在大众的领域而不继续控制，它可能会被用于商业应用程序中。

9.3.2 引用软件发行者

大众可得的软件是一个发行本，你所使用的免费软件必须考虑到你研究中使用的其他版本。一些免费软件有相应的信息以便引用，而另一些则没有。引用软件发行者的因特网 URL 来代替杂志/卷册部分通常是可行。例如：

Felsenstein，J.1993. PHYLIP(Phylogeny Inference Package) version 3.5c. Distributed by the author at ftp://evolution.genetics.washington.edu/.Department of Genetics，University of Washington，Seattle.

Gilbert，D. G.，1996. SeqPup，biosequence editor & analysis platform，version 0.6. Bionet.Software, July 1996. 〈news://4rb7hr$6rc@usenet.ucs.indiana.edu〉 See also ftp://iubio.bio.indiana.edu/molbio/seqpup/。

其中一些可获得的程序已在相同位置时间达 10 年左右，所以带有因特网定位暂时应该是没有问题的。

致谢

许多生物科学的免费软件的开发者(其中的一些是在此提到的)是该文件的真正作者。如果你使用他们的软件，请让他们知道你觉得它很有用。通常一个程序

建立在其他程序的基础上。我想要感谢的有：Jonathan Kans，Joseph Felsenstein，Michael Zuker，Gary Olsen，Dan Davison，Rob Harper，Dave Kristofferson，Reinhard Doelz，Rainer Fuchs，Peter Markiewicz，Thure Etzold，Xiaoqiu Huang，Des Higgins，Harry Mangalam，Jim Brown，Bill Pearson 和许多对此有贡献的人。提出建议、批评和软件应运行方式的观点的用户也为使免费软件更利于大家使用而做出了许多贡献。20 世纪 80 年代以前在 Intelligenetics 的 GenBank 主页为免费分子生物学软件开办了一个档案馆，我们为他们先行的努力而表示感谢。

附录：软件列表

这 150 多个在分子生物学和相关领域的免费软件程序的列表，虽然从各方面看不很详尽，但包含的许多可用于 Mac 和/或 MS Windows 计算机。

计算机操作系统关键词：M—MacOS，W—MS Windows 或 MS DOS，O—Other(通常是 UNIX)，一些软件文档缩写：

ebi — ftp://ftp.ebi.ac.uk/pub/software/ 或 http://www.ebi.ac.uk/software/software.html

iubio — ftp://iubio.bio.indiana.edu/molbio/ 或 http://iubio.bio.indiana.edu/soft/molbio/

IUBio 其他的分子生物学站点有：

ftp://ftp.funet.fi/pub/sci/molbio/iubiomolbio

ftp://ftp.sunet.se/pub/molbio

ftp://ftp.nig.ac.jp/pub/mirror/IUBIO/molbio

ftp://ftp.uam.es/pub/mirror/molbio

ftp://ftp.pasteur.fr/pub/GenSoft/mirrors/IUBIO/molbio

http, ftp://mic3.hensa.ac.uk/hosts/iubio.bio.indiana.edu/molbio/

ftp://bioinformatics.weizmann.ac.il/pub/software/mac and software/ibmpc

一些电邮地址和主页 URLs 可能过时了，除非特别声明，所有列出的软件都由作者授权，且可免费用于非商业使用，特别授权部分应该被标明。一些软件是共享软件，作者需要使用费。

ABACUS — M，W，O

ABaCUS 是一个非装饰程序，它用来研究外显子和蛋白质结构亚基之间的假定的对应关系。

作者：Arlin Stoltzfus，arlin@is.dal.ca

文档：iubio/evolve/abacus/

ADE-4 M

ADE-4 是一个为 Mac 机设计的多元分析和图像显示软件包。

作者：Olivier J. M et al., Jean-Michel.Olivier@biomserv.univ-lyon1.fr

主页：ftp://biom3.univ-lyon1.fr/pub/mac/ADE/ADE4,
http://biomserv.univ-lyon1.fr/ADE-4.html

AMPLIFY M

这个 Macintosh 软件是用来设计、分析和激发涉及聚合酶链式反应(PCR)实验的。Amplify 会查找附近的匹配的目标序列并显示使用不同引物的结果，它能为匹配的序列和内部重复检查 oligos。

作者：Bill Engels，WREngels@macc.wisc.edu

文档：iubio/mac/amplify*, ebi/mac/

ANALYZESIGNALASE M

基于 Macintosh 系统，运用 von Heijne 算法对于哺乳动物的信号序列进行分析和预测。用重量矩阵方法来预测分泌肽中信号肽酶切割信号肽的位点。

作者：Ned Mantei

文档：iubio/mac/analyze-signalase*

ANCESTOR W

Ancestor 设计用来从一系列系统发生关系已知的同源氨基酸序列推断祖先的氨基酸序列。

作者：Jianzhi Zhang，zhang@imeg.bio.psu.edu

文档：iubio/ibmpc/ancestor*

ANNHYB W

Annhyb 是用于 Windows 95 的小程序，它能计算不同 DNA 序列参数。

作者：O. Friard, G. Stefanuto，friard@ba.cnr.it

主页：http://area.ba.cnr.it/~e105of01/annhyb221.zip+

ANTHEPROT W，O

ANalyze THE PROTeins(ANTHEPROT)是一个软件包，它包含对物理化学特性(疏水性、阻力、弹力、溶解性、两亲性)的研究，第二结构预测，Chou 和 Fasman、

Garnier、Gibrat、Deleage、Levin，对转膜区域和结构领域的预测，多序列比对，使用 PROSITE 和 PATMAT 查找生物位点、使用 FASTA 查找同源蛋白质，几个序列之间的同源性比较，以及查看和把握来自 PDB 的蛋白质结构。

作者：G. Deleage，deleage@ibcp.fr & C.Geourjon，geourjon@ibcp.fr

主页：http,ftp://www.ibcp.fr/

文档：iubio/ibmpc/antheprot*

AUTOMATIC-BLAST　　M

这是一个 AppleScript，它自动在每日或每周指定的时间，通过 e-mail 把序列上传到在 ncbi.nlm.nih.gov 的 BLAST 服务器上。这个脚本使用了 Eudora 程序。

作者：Brian Osborne，bosborne@nature.berkeley.edu

主页：http://pgebaker4.pw.usda.-gov/bio/bio.html

文档：iubio/mac/automatic-blast.*

BCM SEARCH LAUNCHER　　M，W，O(PERL)

BCM Search Launcher 是组织分子生物学相关搜索的网页和在网页按功能可得的分析服务的合成，它为相关搜索提供输入框。UNIX 和 Mac 机有一个批处理用户接口，它允许输入序列作为背景任务而被自动查寻，返回的结果做单独的 HTML 文档。需要 Perl 支持。

作者：Randall F. Smith et al.

主页：http://gc.bcm.tmc.edu:8088/search-launcher/launcher.html

BUFFERSTACK　　M

提供适当的信息，BufferStack 将会在使用温度时，根据指定的 pH 和离子强度构建一个完全缓冲液的制法。

作者：Rob Beynon

文档：iubio/mac/buffer*，ebi/mac/bufstack*

CABUFFER　　W

这个程序允许计算在一个混合物中的所有离子浓度，这些缓冲液，如 EDTA、EGTA、NTA、HEDTA、柠檬酸盐、Ca 结合蛋白等。校正温度、离子强度和 pH。

作者：Jochen Kleinschmidt，kleinschmidt@mcclb0.med.nyu.edu

文档：iubio/ibmpc/cabuf*

CAIC　　M

CAIC(Comparative Analysis by Independent Contrasts)是计算比较数据的系统发生独立对比，允许有效的统计学上的改编的假说测试。

作者：Andy Purvis, Andrew Rambaut, Andrew.Rambaut@zoology.ox.ac.uk

主页：http://evolve.zps.ox.ac.uk/CAIC/CAIC.html

CAP　　M，W，O

Contig Assembly Program(CAP)敏锐的检测片段重叠。C程序和命令行程序能用于 SeqPup 及其他。

作者：Xiaoqiu Huang，huang@cs.mtu.edu

主页：ftp://cs.mtu.edu/pub/huang

文档：iubio/align/cap*

CGR　　M

一个用 Chaos Game Representation 提供核苷酸序列数据的面向超级文本的软件。

作者：Heikki Lehva，lehvaslaiho@cc.helsinki.fi

文档：iubio/mac/cgr*，ebi/mac/cgr*

CLUSTAL W　　M，W，O

一个多序列比对程序，用途广。Nucleic Acids Res.22，4673-4680(1994)。

作者：D. Higgins et al.

主页：ftp://ftp-igdmc.u-strasbg.fr/pub/clustal*

文档：ebi/mac/clustalw*，ebi/dos/clustalw*，iubio/align/clustal*

CLUSTAL X　　M，W，O

CLUSTAL X 是 CLUSTAL W 序列比对程序的一个图形界面。序列比对显示在一个窗口，还有下拉菜单。

作者：Thompson J. D et al.，julie@igbmc.u-strasbg.fr

主页：ftp://ftp-igbmc.u-strasbg.fr/pub/ClustalX

CODON FREQUENCY ANALYZER　　W

本程序帮你通过比较已知 DNA 编码区密码子频率来鉴别 DNA 的编码区，适

用于所有生物。

作者：Ballyclaire Analysis

文档：iubio/ibmpc/codon*

CODONBIASINDEX M

这个 codon bias index 是由 Bennetzen 和 Hall 创建的统计数据，定量广泛使用的密码子和不偏好的密码子。

作者：Tom Ritch，ritch@seas.ucla.edu

文档：iubio/mac/codonbioasindex*

COMAP W

从酶切数据库构建小 DNA 片段限制酶切图谱，该程序有图形用户界面。

作者：Kay Hofmann，khofmann@cipvax.biolan.uni-koeln.de

文档：iubio/ibmpc/codon*，ebi/dos/

CONSINSPECTOR M，W，O

ConsInspector 用一个扩展的重量矩阵预编译程序库来描述转录因子结合位点，以扫描与这些位点匹配的核酸序列。

作者：K. Frech et al. frech@gsf.de

主页：ftp://ariane.gsf.de/pub/

COVARIATION M

是一个排列的 RNA 序列系统发生比较分析的面向超级文本的软件。

作者：James W. Brown，jwbrown@mbio.ncsu.edu

文档：iubio/mac/covariation*，ebi/mac

CPRIMER M

CPrimer 评估寡核苷酸是否可作为 PCR 引物。它显示退火温度、干涉结构，也能搜索最佳扩增引物对。

作者：Greg Bristol，gbristol@ucla.edu

文档：iubio/mac/cprimer*

DCSE W，O

Dedicated Comparative Sequence Editor(DCSE)是一种多序列编辑器。它能用于编辑蛋白质、DNA 或 RNA 序列，分子结构能被结合于序列。它提供许多特性，

如字符和结构的彩色显示，自动与已排列的序列比较、序列分组、序列或类型搜索、标准系统、结合的 RNA 结构检查、在线超文本帮助等。

作者：Peter De Rijk，derijkp@reks.uia.ac.be

主页：http://www-rrna.uia.ac.be/~peter/dcse

文档：ebi/dos/dcse

DIGEST W

Digest 扫描 DNA 序列文件以查找酶切位点，它提示用户到指定的酶切位点。如果它们在酶切位点数据库，它写出所有的位点，并按长度分类。

作者：Ramin Nakisa，ramin@ic.ac.uk

文档：iubio/ibmpc/digest*，ebi/dos/digest*

DIGISPEAK M

在 Graf-Bar 或相似的音波数字转换器的帮助下以读取测序胶。

作者：Ned Mantei，bcmantei@aeolus.vmsmail.ethz.ch

文档：iubio/mac/digispeak*，ebi/mac/

DISPAN W

DISPAN(genetic DIStance and Phylogenetic ANalysis)用来计算：均值杂合现象和总体的标准误、基因多样性和相关的参数、标准遗传距离和错误、总体的 DA 距离。它也能进行进化树分析和引导程序测试。

作者：Tatsuya Ota，imeg@psuvm.psu.edu

文档：iubio/ibmpc/dispan*

DNA RUNS M

DNA Runs 是 DNA 序列多态性和分散数据的重要性测试。

作者：John H. McDonald，mcdonald@udel.edu

主页：http://udel.edu/~mcdonald/

文档：iubio/mac/dna-runs*

DNA SLIDER M

DNA Slider 是 DNA 序列数据多态位点与固定的差别的比率的异质(源)性的重要性测试。

作者：John H. McDonald，mcdonald@udel.edu

主页：http://udel.edu/~mcdonald

文档：iubio/mac/dna-slider*

DNA STACKS M

DNA Stacks 是一个 HyperCard 堆栈软件包，它为浏览和处理分子数据提供有用性。DNA Translator 包括基因作图工具，绘制并显示两个线性化的基因图以作比较。Aligner 是一个多序列编辑和显示的堆栈。密码子选择显示各种生物和细胞器的密码子和氨基酸选择数据。

作者：D. J. Eernisse，DEernisse@fullerton.edu

主页：http://biology.fullerton.edu/people/faculty/doug-eernisse

文档：iubio/mac/dnastacks*，ebi/mac/

DNA WORKBENCH M，W，O(PERL)

是一个序列搜索和操作的程序，它提供在 GenBank 和其他数据库有力的快速搜索，客户-服务器访问远程数据库和程序。它的许多序列操作功能包括反向互补、显示读框、DNA-蛋白质的互换、编辑、查找酶切位点、查找人重复序列、与数据库或用户文件比较序列、查找一段序列的规则表达等，它需要 Perl 支持。

作者：James Tisdall，tisdall@cbil.humgen.upenn.edu

主页：ftp://cbil.humgen.upenn.edu/pub/dnaworkbench

DNADRAW M

DNAdraw 用来制备用于发表的 DNA 和蛋白质序列。

作者：Marvin Shapiro，mbs@kias.com

文档：iubio/mac/dnadraw*

DNAFRAG W

用于胶中 DNA 限制酶图谱或蛋白质大小。如果标准分子质量在同一胶中跑，它能计算酶切片段或肽段的大小，通过它们的迁移率可以画出标准分子质量的曲线。

作者：John Nash，Nash@biologysx.lan.nrc.ca

文档：iubio/ibmpc/dnafrag*，ebi/dos/dfrag*

DNASP W

DNASP 是一个 Windows 软件包，对 DNA 数据，如成千上万个碱基的多序列

执行多方面的群体遗传学分析。

作者：Julio Rozas & Ricardo Rozas，julio@porthos.bio.ub.es

主页：http://www.bio.ub.es/~julio/DnaSP.html

文档：ebi/dos/dnasp

DOTPLOT　W

MS-DOS 的 dotplot 程序。

作者：Ramin Nakisa，ramin@ic.ac.uk

文档：iubio/ibmpc/dotplot*, dprel3*, ebi/dos/dotplot*

DOTTY PLOTTER　M

Dotty Plotter 是一个分子生物学方面画点矩阵序列对比工具。Dot plots 是用浏览两个核苷酸序列或蛋白质序列的同源性区域的程序。

作者：D. Gilbert，software@bio.indiana.edu

主页：iubio/mac/dottyplot*, ebi/mac/

DOUBLE DIGESTER　M，O

该程序是用于帮助分子生物学研究人员汇集 DNA 双酶切实验的酶切图谱。

作者：L. Wright，wright-lawrence@yale.edu

文档：iubio/restrict-enz/, ebi/mac/

DPRIMER　M

在 Macintosh 系统中用于计算简并引物 T_m 的程序。

作者：Haoyuan Chen，hchen@bimcore.emory.edu

文档：iubio/mac/dprimer*

EDITVIEW　M

ABI Sequencer 的 DNA 浏览器。它是一个允许查看和打印已分析的来自 ABI PRISM Genetic Analyzer 包含序列数据的样品文件。

作者：EditView@perkin-elmer.com

主页：ftp://ftp.abd.perkin-elmer.com/pub/public/Sequencing/EditView/EditView1.0.1.sea.hqx

文档：iubio/mac/editview*

ENTREZ　　M，W，O

Entrez 是 NCBI 的分子序列检索系统。它提供一个整合方法来获得访问核酸和蛋白序列信息，访问与发表序列相应 MEDLINE 引文和与序列相关的 MEDLINE 子集。

作者：various at NCBI

主页：ftp://ncbi.nlm.nih.gov/entrez/

ENZYME KINETICS　　M

Enzyme Kinetics 是 Mac 机的面向超级文本的软件。它计算并且做酶催化反应的生物化学动力学图。

作者：D. Gilbert，software@bio.indiana.edu

文档：iubio/mac/enzymekinetic*，ebi/mac/enzymekin*

ESEE　　W

Eyeball SEquence Editor(ESEE)是 MS DOS 程序。

作者：Eric L. Cabot，cabot@gcg.com

文档：iubio/ibmpc/esee*

FASTA　　M，W，O

FASTA 序列比对程序是由 FASTP 改进而来，最初是在 Science[Lipman and Pearson，(1985) Science 227，1435~1441]上描述。

作者：Bill Pearson，wrp@virginia.edu

主页：ftp://ftp.virginia.edu/pub/fasta/

文档：iubio/search/fasta*

FASTDNAML　　M，W，O

FastDNAml 是 Joseph Felsenstein 的 DNAML(PHYLIP 的一部分)的较快版本，用户在使用前应查阅 DNAML 的说明文件。

作者：Gary J. Olsen et al.，gary@phylo.life.uiuc.edu

文档：iubio/evolve/fastdna*

FOLDIT　　M

FoldIt 是一个分子模型程序，它形象化的操纵蛋白质，能分析 1600 个残基的蛋白质，并能提取许多结构特性：Ramachandran plots、二硫键、氢键，以及还有关于原子的参数统计。

作者：Jean-Claude Jesior，jean-claude.jesior@imag.fr

主页：ftp://ftp.imag.fr/pub/TIMC/FoldIt.html

文档：iubio/mac/foldit*, ebi/mac/

GCUA　M, O

General Codon Usage Analysis(GCUA)用来计算可能与一系列基因的访问密码子选择模式有关的各种参数。用户能在数据表里查看所有密码子选择(或任何其他的统计)或者单个基因。该程序的特性包括：密码子选择的多变分析(RSCU)和氨基酸模式、计算密码子选择的频率、RSCU 值、氨基酸频率数据、碱基组成、基因距离，并能分析整个复杂原核基因组。

作者：James O. McInerney，J.mcinerney@nhm.ac.uk

主页：ftp://ftp.nhm.ac.uk/pub/gcua/

文档：iubio/mac/gcua*

GEL　M, W

一个用于计算琼脂糖凝胶中 DNA 片段大小的应用程序。

作者：Jean-Michel Lacroix，lacroix@medac.med.utoronto.ca

文档：iubio/mac/gel-jml, iubio/ibmpc/gel-jml, ebi/mac/gel-jml, ebi/dos/gel-jml

GEL　W

GEL 采用一系列标准 DNA 片段大小和迁移率并且预测未知片段大小，用最小的方区以适合迁移率和片段大小的关系。

作者：John R. Thompson

文档：iubio/ibmpc/gel*, ebi/dos/gel/

GEL FRAG SIZER　M

Gel Frag Sizer 是以迁移率来计算酶切片段大小的面向超级文本的软件。提供两种估算大小的方法：Elder 和 Southern 的局部交替法或立方拼接法。

作者：D. Gilbert，software@bio.indiana.edu

主页：iubio/mac/gelfragsizer.*

文档：ebi/mac/

GEL MANAGER　W

Gel Manager 是一个友好的运行于 MS Windows 的用户程序，它包括图像处理技术和数据分析选项。它能处理不同种的数据，如 RFLP、RAMM、RAPD、微小

卫星和其他指纹技术。它对研究遗传关系、分类学、分类、流行病学等有很大帮助。

作者：Carlos Vaquerizo，Joaquin Dopazo，dopazo@samba.cnb.uam.es

主页：ftp://ftp.cnb.uam.es/software/molbiol/gel_man

GENEDOC　W

GeneDoc 是一个全特征多序列比对编辑程序和描影法效用。它会帮助你将你的遗传研究工作发表(它能提供阴影、页码、字体显示特征)。

作者：Karl Nicholas，ketchup@cris.com

主页：http://www.cris.com/~ketchup/genedoc.shtml

文档：iubio/ibmpc/genedoc*

GENEMASTER　W

是一个小的执行序列搜索的软件包，查找富含 GC 区，用于各种各样的起始密码子和编码区的翻译、酶切分析。

作者：Shawn Abigail，ad873@freenet.carleton.ca

文档：iubio/ibmpc/genemast*

GENETREE　M，W

GeneTree 能计算物种树中包埋基因树的代价，真实地显示定位和基因重复数量以及丢失，查找最佳的物种树。

作者：Roderic D. M. Page，r.page@bio.gla.ac.uk

主页：http://taxonomy.zoology.gla.ac.uk/rod/genetree/

GEPASI　W

Gepasi 用于模拟化学和生化反应动力学系统，它能模仿在几个不能反应容器中的无规卷曲和反应进程。用该程序能将其结果绘制成二维或三维图形。无规卷曲能用于代谢控制分析和线性稳定性分析。

作者：Pedro Mendes，prm@aber.ac.uk

主页：http://gepasi.dbs.aber.ac.uk/softw/gepasi.html

文档：ebi/dos/

HDPROBE　M

HDProbe 接受输入的探针和等位基因序列，然后显示关于它们在异型双链分子中定向的序列。

作者：Marvin Shapiro，mbs@pa.net

主页：iubio/mac/hdprobe*

HELIXVU W

HelixVu 图解一个 80bp 的 DNA 区作为一个带有序列的清单打印在螺旋图上面。这对于展示 DNA 修饰间的空间关系很有用处。

作者：Richard Seyler

文档：iubio/ibmpc/helixvu*

HYPER W

Hyper 是 MS Windows 下的一个分析酶动力数据的程序。酶动力数据属于非线性回归，其结果以 5 种标准图形表格方式显示并且打印下来。

作者：J S Eastery，jse@liverpool.ac.uk

文档：iubio/ibmpc/hyper*

HYPERPCR M

用 Rychlik 算法来计算 PCR 反应的最佳退火温度。

作者：Brian Osborne，bosborne@violet.berkeley.edu

文档：iubio/mac/hyperpcr*，ebi/mac/

INTRON ANALYZER W

动植物的内含子的碱基组成有基本的不同，该程序检测内含子并查找规律性。对于一个给定的内含子可以在 5′或 3′端比较它们，或研究其与外显子的相接部分。本程序将建立一个一致序列以显示在每个位置上的最频繁的碱基，包括一个图形表。

作者：Michael Liss，LISS@alf1.ngate.uni-regensburg.de

文档：iubio/ibmpc/intron-analyzer*，ebi/dos/intana*

LALNVIEW M，W，O

LalnView 是一个为可视局部两序列比对的图形程序。序列由彩色矩形描绘全部同源性图形。它能显示序列特征(活性位点、结构域、基序、前肽等)。LalnView 是一个分析成对的比对并且能使序列同源和其结构与功能相连接的好工具。

作者：Laurent Duret，duret@dim.hcuge.ch

主页：ftp://expasy.hcuge.ch/pub/lalnview

LINES&KINETICS M

以图形方法用正常或对数数据来计算线性回归，微生物培养的倍增时间和酶

反应的动力学参数。

作者：Manuel G. Claros，claros@uma.es

主页：http://www.ie.embnet.org/embnet.news/vol5_1/kinetics.html

文档：iubio/mac/lines-kinetics*

LINKAGE-1 M

设计 Linkage-1 是用来帮助遗传学家来检测和估计分离的后代间的连锁。

作者：Karl A. Suiter，ksuiter@acpub.duke.edu

主页：iubio/mac/linkagel*

LINTR W，O

这些程序是用来检测给定的系统发育树的生物钟和运用核苷酸或氨基酸序列构建线性化树。

作者：Naoko Takezaki，ntakezak@lab.nig.ac.jp

文档：iubio/evolve/lintr/

LOOPDLOOP M，W，O(JAVA)

Loopdloop 是分子生物学上用来绘画和编辑 RNA 二级结构的一种工具，它有 MacOS-specific 和 Java 两种版本。选项 Mulfold 可以产生 RNA 折叠，并显示出来。Loopviewer 是一种相关的程序，虽然缺少编辑功能，但应用简单。

作者：D. Gilber，software@bio.indiana.edu

主页：iubio/loopdloop/

文档：ebi/mac/loop*

MACAW M，W

MACAW 是一种关于定位、分析和编辑在众多复杂的序列中相似的序列，并且进行多重比对。它包括序列搜索、编辑和展示。它是一种非常好的程序，可以让我们找到同源性的序列。

作者：Greg Schuler, Stephen Altschul，schuler@ncbi.nlm.nih.gov

主页：ftp://ncbi.nlm.nih.gov/pub/macaw

文档：iubio/ncbi/macaw/，ebi/dos/macaw*

MACBOXSHADE M，O

这种程序可以做出蛋白质、DNA 比较好的打印效果，它不是通过自身来进行

比较，而是用多重比较的程序。可以以 Postscript、EPSF、PICT、RTF. 或 ASCII 的形式输出。在多重比较里不同和相似的会以不同的颜色和底纹加以突出，这里有许多关于底纹、序列号、共有区输出等选项。

作者：Michael D. Baron，michael.baron@bbsrc.ac.uk(macos)，Kay Hofmann(original)

主页：ftp://ulrec3.unil.ch/pub/boxshade/macboxshade

文档：iubio/mac/macboxshade*

MACPATTERN　　M

Macpattern 是 Macintosh 系统应用于蛋白质模型搜索(用 PROSITE)和区组概要搜索(用 BLOCKS)。Macpattern 通过 PROSITE 数据库有助于发现支持新蛋白质序列假定功能的模型。区组概要搜索用 BLOCKS 数据库和统计分析(最大片段分析和 Eguchi-Seto 方法)。

作者：Rainer Fuchs，rainer_fuchs@glaxo.com

文档：iubio/mac/macpattern*，ebi/mac

MACPLASMAP　　M

如果你研究相关的环形质粒图形，你就会发现 MacPlasmap 是一种必不可少的工具。它可以绘画、储存、打印高质量的含有你所定义的数据的质粒图形。

作者：Jingdong Liu

文档：iubio/mac/macplasmap*，ebi/mac/

MACPROT　　M

MacProt 分析蛋白质二级结构、链的柔性、水疗法、螺旋轮等。

作者：Peter Markiewicz

文档：iubio/mac/plota/，ebi/mac/plota_*

MACSTRIPE　　M

MacStripe 是一个推测和分析蛋白质序列中潜在的卷曲螺旋。MacStripe 对于任何想研究蛋白质中潜在卷曲螺旋的人来说都是一个必不可少的理想工具。通过一个全 Macintosh 界面，分析结果(粗略的数据和具有出版质量的图表)都可以轻松的输出到其他的软件上。MacStripe 用 Andrei Lupas's COILS2 算法来计算详细信息和可信赖的卷曲螺旋假设。

作者：Alex Knight，aek4@york.ac.uk

主页：http://www.york.ac.uk/depts/biol/units/coils/coilcoil.

html
文档：iubio/mac/macstripe*

MACT　M

MACT 是 Macintosh 的一组用矩阵的方法来构建和评估来源于氨基酸序列的无根树形结构。

作者：Angela Luettke, Rainer Fuchs

文档：ebi/mac/mact_*，iubio/mac/mact*

MAP MANAGER　M，W

Map Manager 是一组通过与显性标记杂交、回交或重组天然菌株来分析遗传实验结果的程序。这是一个具体的数据库程序，它可以轻松的储存、检索和显示这些实验图的信息，同时，它也有搜索工具而且可以分析实验结果。这些工具有助于使用者决定其位置和位置次序。

作者：Kenneth F. Manly，kmanly@mcbio.med.buffalo.edu

主页：http，ftp://mcbio.med.buffalo.edu/

文档：iubio/mac/map-manager*，ebi/mac/mapmanager*

MAPMAKER　M，W，O

MapMaker 是一种联系分析模式，是用来帮助构建实验中多种标记的主要联系图，也可分析显性、隐性、共显性标记的多点式联系。

作者：Whitehead Institute for Biomedical Research，mapmaker@genome.wi.mit.edu

主页：ftp://genome.wi.mit.edu/distribution/mapmaker3

文档：iubio/mapmaker/

MATERIALS & METHODS　M

一种存取实验室程序的系统，此系统可重复存取分子生物学常用程序。

作者：James W. Brown，jwbrown@mbio.ncsu.edu

文档：iubio/mac/mandm*，ebi/mac/matmeth*

MATILDA　W

一种独特 DNA 数据库管理系统，能帮助我们从大量的序列和基因信息中获得重组 DNA 所需的高水平信息，它是功能性数据和图式数据的集中表现。序列和功能性数据可以从序列档案和附加信息获得。当重组 DNA 克隆完成后，它们

的描述被附加在数据库中以备后用。

作者：B. Isralewitz，D. Shalloway

文档：iubio/ibmpc/matilda*

MATIND AND MATINSPECTOR　　M，W，O

MatInd 是一种从 IUPAC 编码确定的短序列中获得同一性矩阵描述的简单而有效的方法。MatInspector 是一种程序，用大量的预先确定的转录因子矩阵描述来定位核苷酸非限制性长度序列的一致性。它指派一个质量比率来匹配，因此允许质量基础滤过和匹配选择。

作者：K. Quandt et al.，quandt@gsf.de

主页：ftp://ariane.gsf.de/pub/

文档：ebi/mac/matind*，ebi/dos/matind*

MEMSAT　　W，O

膜蛋白结构和拓扑学。

作者：David T. Jones，jones@bsm.bioc.ucl.ac.uk

主页：ftp://ftp.biochem.ucl.ac.uk/pub/MEMSAT

METREE　　W

一种推断和检察最小进化树的系统，即为了一些序列试图去寻找那些有最少枝条长度总量的最小进化树，以便确定我树、他树的显著不同并将之打印。

作者：Andrey Rzhetsky, Masatoshi Nei，aurl@psuvm.psu.edu

文档：iubio/ibmpc/metree*

MITOPROT　　M，O

提供一系列的参数帮助我们对线性靶序列的理论估计和可输入性。MitoProtII 为我们预测线性蛋白靶序列和叶绿体蛋白提供可能。

作者：Manuel G. Claros，claros@uma.es，Pierre Vincens

主页：ftp://ftp.ens.fr/pub/molbio/，ftp://ftp.rediris.es/software/incoming/science/

文档：iubio/mac/mitprot*,ebi/mac/

MOLWT　　W

该程序通过内部化学方程式计算分子质量，并且通过内部化学方程式和浓度单位给予相应的单位。

作者：John A. Kiernan，jkiernan@julian.uwo.ca

文档：iubio/ibmpc/molwt*，ebi/dos/molwt*

MFOLD(MAC) M

通过最小自由能来推断 RNA 的二级结构，包括依靠温度拿不准的折叠，见 PCFold。

作者：D. Gilbert(mac port)，M. Zuker(MFold)

主页：iubio/mac/mulfold*，ftp://snark. wustl. edu/pub/(MFold)

文档：ebi/mac/

NIH IMAGE M

Image 用来测量使用者兴趣限定区域的范围、平均浓度、重心和角度。也可执行自动颗粒分析和测量路径长度和角度。

作者：Wayne Rasband

主页：ftp://zippy.nimh.nih.gov/pub/nih-image/，http://rsb.info.nih.gov/nih-image/

NJBAFD W，O

这些程序用于从微卫星 DNA 或其他基因标记的等位基因频率构建一个邻域连接或 UPGMA 树，并能计算杂合现象和 Gst。

作者：Naoko Takezaki，ntakezak@lab.nig.ac.jp

文档：iubio/evolve/njbafd/

NJPLOT M，W，O

NJPlot 是一个画系统树的程序，它能通过嵌套的刮弧的方法(如 PHYLIP 构建树)处理文件描述树。特点：图形界面允许你随时随地重起一个树和交叉分支。引导程序值在邻近于内部分支显示，分支长度能随意地显示，树图能存成 PostScript 或 PICT 文件。

作者：Manolo Gouy，mgouy@biomserv.univ-lyon1.fr

主页：ftp://biom3.univ-lyon1.fr/pub/mol_phylogeny/njplot

文档：ebi/mac/

NONCODE W

读取一个包含核苷酸序列的 ESEE 文件，用 Kimura 2 参数据模式产生一个距离矩阵。

作者：Eric L. Cabot, cabot@gcg.com

文档：iubio/ibmpc/noncode*

NUMCLONE W

估计从基因组文库中筛选的克隆数目，以便找到期望的克隆。

作者：John Nash，Nash@biologysx.lan.nrc.ca

文档：iubio/ibmpc/numclone*

OLIGOBASE W

这是一个 Windows 的共享软件，用来组织和分类生物图书馆的寡核苷酸收集。它存储寡核苷酸信息，根据物种标准选择子集，以表格形式按顺序打印，计算分子质量和 T_m，对寡核苷酸进行操作。

作者：Igor Sidorenkov，sidorenk@rocketmail.com

主页：http://lochfort.com/oligobase

文档：iubio/ibmpc/obase*

OLIGOCR M

OligoCR 是一个组织和分类寡核苷酸收集的管理工具。能存储 oligo 的信息，如目录(PCR，测序)、它的应用、它的描述。只要点击鼠标就能完成搜索自己的 oligo 数据库。具有校正序列的能力。

作者：Yongming Sun，ysun@hdklab.wustl.edu

主页：http://hdklab.wustl.edu/~ysun

文档：iubio/mac/oligocr*

OLIGOMUTANTMAKER W

OligoMutantMaker 能简化设计和筛选寡核苷酸介导的单氨基酸代替实验，这通过搜索导入一个酶切识别序列到密码子突变的替代位置核苷酸序列。

作者：Kevin Beadles et al.

文档：iubio/ibmpc/oligo*, ebi/dos/oligo*

ONIX W

MS Windows 程序，Onix 允许用户用 PDB 的已知三维结构检查蛋白质。设计用于在蛋白质中进行配体结合位点的结构研究。Onix 是具有交互式高性能界面的软件，快速三维分子图形和亲水界面的分析。

作者：A. S. Ivanov et al.，ivanov@ibmh.msk su

主页：ftp://org.chem.msu.su/pub/software/onix/

P1 CLONES M

保持构建自噬菌体 P1 克隆系统跟踪的简单面向超级文本的软件。

作者：Ken Abremski，sabremske@esvax.dnet.dupont.com

文档：iubio/mac/plclones*

PAML M，W，O

Phylogentic Analysis by Maximum Likelihood(PAML) 包括三个模型装配的主程序，用核苷酸或氨基酸序列数据进行系统树构建。

作者：Ziheng Yang，z.yang@ucl.ac.uk

主页：ftp://abacus.gene.ucl.ac.uk/pub/paml

文档：iubio/evolve/paml*

PCFOLD W

Michael Zuker 的 RNA 折叠程序，它是用能量最小算法来预测 RNA 结构中的茎环区。参见 MulFold 和主页下载的 UNIX 版本。

作者：Michael Zuker et al.，zuker@snark.wustl.edu

主页：ftp://snark.wustl.edu/pub/

文档：iubio/ibmpc/pcfold*，ebi/dos/pcfold

PHYLIP M，W，O

PHYLogeny Inference Package(PHYLIP)是系统树分析的多个程序，包括简约性、相容性、距离矩阵常量(“进化简约性”)以及各种各样可能的数据。

作者：Joseph Felsenstein，joe@genetics.washington.edu

主页：http，ftp://evolution.genetics.washington.edu/

文档：iubio/evolve/phylip*,ftp://ftp.nig.ac.jp/pub/UNIX/phylip，ftp://ftp.bioss.sari.ac.uk/pub/phylogeny/phylip

PHYLODENDRON M，W，O(JAVA)

Phylodendron 是一个绘制系统树的应用程序，它读取 New Hampshire(Newick) 格式的数据，它允许装饰和编辑树。

作者：D. Gilbert

主页：iubio/java/apps/trees/

PHYLTEST W

一个测试系统发生的假设程序，有三个可选的进化系统树的比较，进行平均成对距离的估计等。

作者：Sudhir Kumar, imeg@psuvm.psu.edu

文档：iubio/ibmpc/phyltest*

PLASMID PROCESSOR W

Plasmid Processor 是一个为科学和教育目的的质粒赠送的简单工具。包括环状和线状 DNA，用户定义酶切位点、基因和多克隆位点，以及可以通过插入和缺失片段操作质粒。创建的图可拷贝到剪贴板或存盘以备今后使用，它也支持打印功能。

作者：T. Kivirauma，P. Oikari and J. Saarela，Dept. of Biochemistry and Biotechnology，University of Kuopio，plasmid@uku.fi

主页：http://www.uku.fi/~kiviraum/plasmid/plasmid.html

文档：iubio/ibmpc/plasmid-processor*，ebi/dos/plasmid

PLASMID-MAKER M

它可以绘制线性和环形的质粒图，也可以有不同的宽度以及填充灰色、箭头等。

作者：Kai-Uwe Froehlich，kaifr@uni-tuebingen.de

主页：http://yeamob.pci.chemie.uni-tuebingen.de/Archiv/PlasmidMaker.html

文档：iubio/mac/plasmid-maker*

PRIMERDESIGN W

PrimerDesign 是一种选择 PCR 引物和寡核苷酸探针的 DOS 程序。它可以很好的检查已知的重复序列和独一无二的序列，随后根据这些数据来设计引物。你设计出很多的前提条件，也都是可以的。它可以处理多至 31 500 个碱基对，并显示出其额外的特点：独特的序列，重复的以及限制位点。

作者：Andreas Becker, Joerg Napiwotzki，becker@ps1515.chemie.uni-marburg.de

主页：ftp://ftp.chemie.uni-marburg.de/pub/PrimerDesign

PRIMER M，W，O

Primer 是一种自动选择 PCR 引物的程序。它可以检测退火温度，补充基因组的重复序列，组成引物二聚体和其他标准，退火温度的计算是根据热力学参数。

作者：Steve Lincoln et al.，primer@genome.wi.edu

主页：ftp://genome.wi.mit.edu/pub/software/Primer2.2

文档：iubio/primer/primer-wi*

PRIMER-MASTER W

它可以自动搜索和选择多种不同 PCR 的最佳引物，分析作为引物或杂交探针的寡核苷酸，编辑适合的新核苷酸序列。

作者：Proutski Vitali，Sokur Oleg，proutski@influenza.spb.su

文档：ebi/dos

PRIMERS! M

Primers!是一种设计引物的共享软件，由 Whitehead Institute Primer2 编写。它可以滚动着列出多对引物，从而可以选择出我们所需要的。

作者：Richard Resnick，rjr@applepi.com

主页：http://www.applepi.com

文档：iubio/mac/primers*

PROANAL W

ProAnal 是一种进行蛋白质多重比对的软件，研究在蛋白质或肽家族里其结构与功能，结构与活性的关系。它通过比较氨基酸序列和其活性，并发现在初级结构不同区域的多种理化特征与活性的相关性。

作者：Alexey Eroshkin，eroshkin@vector.nsk.su

文档：iubio/ibmpc/proanal*，ebi/dos/

PROANALYST W

ProAnalyst 是研究由于功能、进化或其他标准而区分开的蛋白质之间的结构不同、结构与活性的关系、搜索基序、蛋白质工程实验以及多种蛋白质功能分析。

作者：Vladimir Ivanisenko，Alexey Eroshkin，eroshkin@vector.nsk.su

文档：iubio/ibmpc/panalyst*，ebi/dos/proanalyst

PROANWIN W

多重序列比对分析蛋白质序列和结构，结构与活性的关系，设计蛋白质工程实验。在三级结构进行比对，寻找线性和空间位点，保守和多变的理化特性，不同理化特征的图表或一组蛋白质序列以及其他的一些功能。

作者：I. Pika et al.，eroshkin@vector.nsk.su

文档：iubio/ibmpc/paw*，ebi/dos/proanwin

PROFILEGRAPH W

一种蛋白质图的分析工具。

作者：Kay Oliver Hofmann，khofmann@biomed.biolan.uni-koeln.de

文档：iubio/ibmpc/prograph*，ebi/dos/pgraph*

PROMFIND W

PromFind 是基于 DOS 的程序，设计 DNA 序列，可以加一些标记来注释假定启动子的位置。

作者：Gordon B. Hutchinson，hutch@netshop.bc.ca

文档：iubio/ibmpc/promfind*

PROMSED W

ProMSED 应用于 Windows，可以自动和手动比较、编辑、分析 DNA 和蛋白质序列。自动比较是根据 Clustal V；手动比较和抽象分析是通过群组分析和氨基酸色彩反应它们的相似，这是比较容易的。

作者：Anatoly Frolov，Alexey Eroshkin，eroshkin@vector.nsk.su

文档：iubio/ibmpc/promsed*，ebi/dos/promsed/

PROPHET W，O

Prophet 是先进的易于使用的软件工具，数据管理和可视化以及统计分析——从简单描述统计到多因素 ANOVA、logistic 回归和非线性的模拟。它也能用于分析生物序列，包括多序列比对、翻译、酶切、蛋白分解的切割分析、PCR 引物设计、BLAST 查询、远程数据库检索等。

作者：Prophet software group，BBN，prophet-info@bbn.com

主页：http://www-prophet.bbn.com/

PROTEIN SEQUENCE ANALYSIS W

该程序是一个氨基酸组成的序列编辑器，可以进行水动力学计算、同位素标记、等电点、UV 光谱、相对疏水性、二级结构预测等。

文档：iubio/ibmpc/prot-sa*

PUZZLE M，W，O

Puzzle 是一个核苷酸、氨基酸或二态数据的最大相似性分析。它可以从分子序列数据重构系统树，并进行大量数据系列的快速树查询。它与 PHYLIP 兼容。

作者：Korbinian Strimmer，Arndt von Haeseler，strimmer@zi.biologie.uni-muenchen.de

主页：ftp://fx.zi.biologie.uni-muenchen.de/pub/puzzle

文档：iubio/evolve/puzzle/, ebi/mac/puzzle, dos/puzzle, UNIX/puzzle

RAMHA　　W

合成的 cDNAs 的随机突变的 Monte Carlo 模拟。

作者：David P. Siderovski，Siderovski@Galen.OCI.UToronto.CA

文档：iubio/ibmpc/ramha*

RASMOL　　M，W，O

RasMol 是蛋白质和核苷酸可视化的分子模型程序。它能读蛋白质数据库(PDB)文件并且交互性的以各种格式提交它们，包括 wire、stick、stick_and _ball、CPK 和 ribbon 等。

作者：R. Sayle，ros@dcs.ed.ac.uk

主页：ftp://ftp.dcs.ed.ac.uk/pub/rasmol/

文档：ftp://kekule.osc.edu/pub/chemistry/software/X-WINDOWS/rasmol*, ebi/mac/rasmol*, software/dos/raswin*

RBINDING　　W

计算结合位点的数量和细胞受体对配体的亲和性(Scatchard 分析)。

作者：Nico van Belzen, Joop van Zoelen，belzen@pal.fgg.eur.nl

文档：iubio/ibmpc/rbindin*

READSEQ　　M，W，O

各种生物序列文件的转换程序。

作者：Don Gilbert，software@bio.indiana.edu

主页：iubio/readseq/

文档：ebi/mac/readseq*

REALIGN　　W

根据肽序列重新排列 DNA 序列，因此改进排列是不太保守的区域。

作者：Rasmus Wernersson，RWer@novo.dk

文档：iubio/ibmpc/realign*

REPFIND W

RepFind(promoter find)是一个MS DOS系统中的鉴定DNA中的普通重复序列的程序。它也能鉴定和模仿载体序列。

作者：Gordon B. Hutchinson，hutch@netshop.bc.ca

文档：iubio/ibmpc/repfind*，ebi/dos/repfind*

RESTDATA W

限制酶数据和系统发生分析，计算一对DNA序列的每个位上的核苷酸替代数目。

作者：Tatsuya Ota，imeg@psuvm.psu.edu

文档：iubio/ibmpc/restdata*

RESTSITE W

用于分析限制酶切位点或片段数据以用于分子系统研究的多个程序集。

作者：Joyce C. Miller

文档：iubio/ibmpc/restsite*

RNA DOTPLOT M

RNA Dotplot是一个打印潜在的RNA序列中碱基对交互作用的点矩阵的简单应用程序。

作者：David S. McPheeters，mcpheeters@biochemistry.cwru.edu

文档：iubio/mac/rna-dotplot*

RNADRAW W

RNAdraw是一个进行RNA二级结构计算的软件结构，可存储矩阵/热力学曲线/概率柱状图。

作者：Ole Matzura，ole@mango.mef.ki.se

主页：ftp://broccoli.mfn.ki.se/pub/rnadraw

RNA_D2 W

RNA_D2是用户友好程序用于使绘图符合审美观和描绘RNA二级结构的重叠构象。此程序使得解螺旋和编辑RNA分子(>1000核苷的长度)简单易操作。

作者：J. Perochon-Dorisse et al.，rnad2@ibcg.biotoul.fr

主页：ftp://hpsrv.biotoul.fr/rna

SAGITTARIUS DNA W

一种以K-tuples为基础揭示外显子/内含子结构的统计包裹。

作者：Victor B. Strelets，strelets@bio.indiana.edu

文档：iubio/ibmpc/sag-exo*，ebi/dos/sag-exo

SAGITTARIUS PIR　W

一种非常简洁方便的原始 PIR 数据库变量，设计的目的是为了帮助使用序列分析数据资源的个人能够在不调动大量内存的情况下进行分析。它包括进行快速的近似查询和选择(姓名、来源、关键字等)或者依据用户的准确定义进行查询。

作者：Victor B. Strelets，strelets@bio.indiana.edu

文档：iubio/ibmpc/sag-pir*

SAGITTARIUS SEQANALREF　W

一种以对话形式进行存储和处理相关信息的程序。这种独特的变量是为了适应经 A. Bairoch 编译的应用程序 SEQANALREF 的数据资源。

作者：Victor B. Strelets，strelets@bio.indiana.edu

文档：iubio/ibmpc/seqanalr*，ebi/dos/sag-sar*

SEND　W

一种用 Nei 和 Jin 的运算法则来计算分析核苷多样性和分歧性错误的程序。

作者：Li Jin

文档：iubio/ibmpc/send*

SENDBS　W，O

一种代替 Nei 和 Jin 的运算法则而分析在种群内部和外部的平均核苷酸的程序，它包括与 Nei 和 Jin 的运算法则不同的带有标准错误的解序列程序。

作者：Naoko Takezaki，ntakezak@lab.nig.ac.jp

文档：iubio/evolve/sendbs*

SEQ-EUDORA-BLAST　M

一种使 BLAST 查询自动化的 Macintosh AppleScript 应用程序。用 Eudora mail 将剥落的 BLAST 服务器片段文件送到 ncbi. nlm. nih. gov。

作者：Brian Osborne, bosborne@nature.berkeley.edu

主页：http://pgebaker4.pw.usda.gov/bio/bio.html

文档：iubio/mac/seq-eudora-blast*

SEQAID II　　W

Seqaid II 是一种分析 DNA 和蛋白质序列的 MS-DOS 程序。功能包括编辑、调整 Needleman-Wunsch 队列、点矩阵比较、片段大小、基本组成、翻译、蛋白结构、限制性位置检索、用密码偏移确定蛋白外显子位置等。

作者：Donald Roufa, D. D. Rhoads

文档：iubio/ibmpc/sequaid*, ebi/dos/sqaid*

SEQAPP　　M

一种 Macintosh 生物序列编辑器、分析员和网络高手(参看 SeqPup)。

作者：D. Gilbert，seqapp@bio.indiana.edu

主页：iubio/seqapp/

SEQPUP　　M，W，O(JAVA)

是一类继承了 SeqApp 进行生物序列分析和编辑的程序。它还包括与网络的链接和外部分析程序。特点包括多样序列整理和编辑，支持多种文件格式编排，序列特点编辑，处理和标记，DNA 和蛋白质的翻译，翻转/补充，远距离方法，形象化的队列描述和带有盒子的序列以及阴影区，网络搜索，外部程序分析，包括 Clustal W 多样队列。CAP 可以在任何支持 Java 的程序上运行，旧的翻译版本可以在 MacOS、MSWin 和 UNIX 上运行。

作者：D. Gilbert，seqpup@bio.indiana.edu

主页：iubio/seqpup/

SEQSIMPRESENTER　　M

SeqSimPresenter 突出排列成一个成阴影棒状——相应的具有一定程度的相似区的序列。它以简洁的方式表现大的序列并使得识别、延伸和分配保留区更加快速。

作者：cbkfr01@mailserv.zdv.uni-tuebingen.de

文档：iubio/mac/seqsimpresent*, ebi/mac/

SEQUIN　　M，W，O

Sequin 是一个由 NCBI 改进的独立的用于递交的个体符合基因库或 EMBL、DDBJ 序列分析的软件工具。它还能够处理简单的带有 MRNA 片段的序列和复杂的带有很长序列的片段，并带有多种注解、系统发生和种群研究。

作者：Jonathan Kans，Colombe Chappey，info@ncbi.nlm.nih.gov

主页：ftp://ncbi.nlm.nih.gov/sequin/, http://www.ncbi.nlm.

nih.gov/sequin

SEQVU M

一种排成一列编辑程序带有分析选项使你工作更快并分析多的序列。它较适合手工修改序列产品使用如 Clustal V 的软件。

作者：James Gardner，j.gardner@garvan.unsw.edu.au

主页：ftp://gimr.garvan.unsw.edu.au/pub/

文档：iubio/mac/seqvu*

SHM W，O

Shm 设计目的是用于帮助分析由免疫球蛋白 B-淋巴细胞引诱的体细胞变异，它以简约性和用单独的躯体突变表现它们从而建立克隆树。

作者：Laurentiu Cocea，cocea@necker.fr

文档：iubio/ibmpc/shm*

SIGMA M，O

一种为集成基因图谱组合(Sigma)而绘成的基因图谱编辑程序，作为一个观察工具，Sigma 将全彩色的基因图谱呈现在观察者面前并方便调整、浏览、控制和打印。它能给使用者整个染色体的透视图，并可以任意地呈现细节图像。其特性可以让使用者找到图谱上的特殊区域，Sigma 使使用者从众多的来源中整理所需的数据。它还配备有一个方便的用户界面使使用者方便进入。

作者：Theoretical Biology and Biophysics Group at LANL，sigma@ncgr.org

主页：http://www.ncgr.org/sigma/home.html

文档：ebi/linkage_and_mapping/SIGMA

SILMUT W

Silmut 帮助你在一个片段中识别可以添加限制性酶切位点和其他由突变片段修改的区域。

作者：Raj Shankarappa，bsh@med.pitt.edu，
K. Vijayananda，vijay@litsun.epfl.ch

文档：iubio/ibmpc/silmut*

SIM2 M，W，O

此程序用于定位排成一列的两条片段，每一条片段都可以长达上千个核苷酸。

作者：Chao K-M et al.，zjing@sunset.nlm.nih.gov

主页：ftp://ncbi.nlm.nih.gov/pub/sim2

SITES M，W，O

SITES 是一种用于分析相似的 DNA 片段的程序。它的基本用途是用于分析复杂的紧密相关的片段数据。

作者：Jody Hey

文档：iubio/evolve/sites/

SIXCUTTERFREQ M

这种面向超级文本的软件包是用于计算 6 种限制性内切酶在少数基因组中的酶切频率，包括 λ 噬菌体、小鼠、小麦、大肠杆菌、酵母和人。这种算法主要的依据是二核苷酸成对的频率。

作者：Brian Osborne，bosborne@nature.berkeley.edu

文档：iubio/mac/sixcutterfreq*

SNEATH ST W

非典型序列比对荧光屏成型片统计程序，是模拟相加常量区使之随机放入不同分子片段并进行比较。

作者：P. H. A. Sneath，mjs@le.ac.uk

文档：iubio/ibmpc/sneathst*

SOLUPRED M，W

这个表格程序使使用者可以预测在大肠杆菌中依据氨基酸含量不同的重组蛋白的溶解性。

作者：Roger Harrison, Dan Diaz，BL275@Cleveland.freenet.edu

文档：iubio/mac/solupred-mac*，ibmpc/solupred*

SORFIND W

Sorfind 是一种 DOS 应用程序，可以添加 DNA 片段文件到一个特殊的表格，从而定位推定的编号外显子。

作者：Gordon B. Hutchinson，hutch@netshop.bc.ca

文档：iubio/ibmpc/sorfind*，ebi/dos/sorfin*

SPECTRUM M，W

Spectrum 是一种 Macintosh 和 MS Windows 程序，用来在 Nexus 格式下读取

系统发生数据，并将这些相应的光谱展示出来。它还可以用来寻找与观测到的光谱相符合的光谱，此程序以 Microsoft Excel 或其他方式输出光谱。

作者：Michael Charleston，Roderic Page，m.a.charleston@bio.gla.ac.uk

主页：http://taxonomy.zoology.gla.ac.uk/mike/spectrum/

SPOMBE-STRAIN M

此程序是面向超级文本的软件包，用来收录 *Schizosaccharomyces pombe* 酵母菌染色剂的遗传型。它适合于 Kai-Uwe Frölichs 酵母菌染色剂。

作者：Doug Drummond，ddrummon@fs2.scg.man.ac.uk

文档：iubio/mac/spombe-strain*

SSU RRNA M

这种面向超级文本的软件包包括整个核糖体数据库工程片段释放 1。这些片段可以通过一系列的系统发生树获得。

作者：James W. Brown，jwbrown@mbio.ncsu.edu

文档：iubio/mac/ssu-rrna*

SWISS-PDBVIEWER M，W

Swiss-Pdb Viewer 是一种可以由 PDB 文件展示的应用程序。可以分析许多蛋白并可以堆积到量纲树空间中。被选种的氨基酸还可以进行分析和比较，可以比较活性区的区别，并测量距离。在原子间扭转角度就像加入或移除氨基酸一样，它具有很多特点。

作者：Nicolas Guex, Manuel Peitsch, ng45767@ggr.co.uk

主页：ftp://expasy.hcuge.ch/pub/PDBViewers/Prot3Dviewer,
http://www.expasy.ch/spdbv/mainpage.html,
http://www.pdb.bnl.gov/expasy/spdbv/mainpage.htm

TACG M，W，O

一种对于限制酶和 DNA 分析的软件。这是一种命令行程序，它可以用在 MacOS，Windows 软件和 Intel 微处理芯片的联盟或其他的，比如 SeqPup。

作者：Harry Mangalam，mangalam@uci.edu

主页：http://hornet.bio.uci.edu/~hjm/projects/tacg/

文档：iubio/restrict-enz/tacg*

TFPGA W

TFPGA(总基因分析工具)是一种 Windows 下的程序，分析异型酶和分子总基因数据。它可以计算简单的描述性的数据、遗传的距离和 F 统计。它也可以做 Hardy-Weinberg 平衡、总的区别和 UPGMA 群集以及 Mantel 检测。

作者：Mark P Miller，mpm2@nauvax.ucc.nau.edu

主页：http://dana.ucc.nau.edu/~mpm2

TOPPRED II M

预言在整个膜蛋白里跨膜片段和假定的拓扑结构。

作者：Claros M. G, von Heijne G，claros@cica.es，gvh@cbs.ki.se

文档：ebi/mac/，iubio/mac/toppred*

TOPS W，O

它可以自动产生和编辑蛋白质布局动画，这些动画是蛋白质二级结构的表达。

作者：Tom Flores，flores@ebi.ac.uk

主页：ebi/pub/contrib/TOPS

TREE DRAW DECK M

一种面向超级文本的软件，可以绘画系统发生树，是 J. Felsenstein 从 Drawgram 和 Drawtree of PHYLIP 演化来的。

作者：D. Gilbert，software@bio.indiana.edu

主页：iubio/mac/treedraw*

文档：ebi/mac/treedraw*

TREECON W

它是构建系统树的软件包。它的优点包括菜单驱动，容易使用的界面，快速处理大的数据库，大的以构建树形结构为基础的测量数据的方法以及矩阵，复杂的选项，比如子集、外围影响检测，还有其他的工具，比如部分比较和信息位置的暗示。

作者：Yves Van de Peer，yvdp@reks.uia.ac.be

主页：ftp://uiam3.uia.ac.be/

文档：iubio/ibmpc/treecon*

TREEVIEW M，W

一种在 MacOS，MS Windows 上绘制系统发生图的程序。它可以读取

NEXUS，PHYLIP，Clustal W 以及相似的树形结构格式。

作者：Roderic D M Page，r.page@bio.gla.ac.uk

主页：http://taxonomy.zoology.gla.ac.uk/rod/treeview.html

文档：iubio/mac/treeview*，iubio/ibmpc/treeview*

VISED　W

一种在 Windows 下编辑/显示序列形象的软件，包括有效的和容易使用的界面。它可以编辑多达 200 个序列 18 000 个碱基，强有力的模型搜索功能；支持多种序列格式；支持从序列库中抽取序列，图表准备，同一性的序列盒式输出、输入 MACAW 比较；在一个或所有的 6 个框架中进行蛋白质序列的推测。

作者：Ken Peters，kperters@qb.island.net

文档：iubio/ibmpc/vised*

VISUAL BLAST AND FASTA　W

这些程序是用来分析 BLAST 和 FASTA 输出结果的相互关系，包括蛋白质序列的比对。它补充了分析工具，即可以自动进行 BLAST 和 FASTA 输出结果的比较，包括多重比较的分析。

作者：Patrick Durand et al.，durand@lmcp.jussieu.fr

主页：http://www.lmcp.jussieu.fr/~durand/

WINDOT　W

一种在 MS Windows 下的点阵图程序。

作者：Ramin Nakisa. ramin@ic.ac.uk

文档：iubio/ibmpc/windot，ebi/dos/windot

WINMGM　W

观察和操纵蛋白质、核酸和有机体的分子，包括 CPK、棒状、球状、带状以及圆柱状的和原子染色的、被选择的原子区域、活性位点以及其他的一些特征。

作者：Mehdi Rahman，Robert Brasseur，mehdirah@fsagx.ac.be

主页：http://www.fsagx.ac.be/info_faculte/info_dep/info_bp/mehdi/winmgm/ winmgmen. htm

WINSEQ　W，O

MS Windows 系统中 ReadSeq 程序可以将各种生物序列文件格式进行转换。

作者：Ramin Nakisa，ramin@ic.ac.uk

文档：iubio/ibmpc/winseq，ebi/dos/winseq，更多信息请查看iubio/readseq/

WPDB　W

WPDB(MS Windows 中的 Protein Data Bank)是 PDB 的简缩版软件包，单结构的查询特点或进行一个强调序列比对和结构重叠的多结构的比较分析。

作者：Ilya N. Shindyalov，Philip E. Bourne，bourne@sdsc.edu

主页：ftp://ftp.sdsc.edu/pub/sdsc/biology/WPDB/

YEAST STRAIN　M

这是一个对酿酒酵母菌株按基因型进行分类的面向超级文本的软件。

作者：Kai-Uwe Froehlich，cbkfr01@mailserv.zdv.uni-tuebingen.de

文档：iubio/mac/yeaststrain*，ebi/mac/

(周传香　欧阳松应　译)

10 用 FASTA3 程序软件包进行灵活的序列相似性搜索

William R. Pearson

10.1 引言

自 15 年前发表第一个快速比较生物学序列方法以来[1]，从最新被克隆了的蛋白质到全部染色体，DNA 和蛋白质序列比较成为生物化学研究的常规步骤。由于 DNA 和蛋白质序列数据库越来越完善，序列相似性搜索更可能揭示与数据库序列的统计学意义上的相似性，并且以此推断与所查询序列的同源性。确实，在与已知功能蛋白质有显著的序列相似性的基础上，甚至在古细菌詹氏甲烷球菌中，40%以上的可读框架的功能都可以被推导出[2]。

本章提供 FASTA 软件包程序的概述(`ftp://ftp.virginia.edu/pub/fasta`)。这里不深入讨论蛋白质和 DNA 序列比较的理论和实践，而集中于更实际的问题，例如，“应该使用哪个 FASTA 程序?”，“应该在统计学意义上使用什么阈值?”，“应该搜索哪个数据库?”，“什么时候应该使用 FASTA 和什么时候应该使用 BLAST?”，还有“什么时候应该改变得分矩阵(scoring matrix)和间隙补偿(gap penalty)?”作为用 BLAST 和 FASTA 进行相似性搜索的优秀综述和局部相似性统计，见参考文献[3]。怎么使用 FASTA 程序鉴定远缘序列的更特定的信息，见参考文献[4]和[5]。对 FASTA3 软件包的统计学估计的详细解释见参考文献[6]。

10.2 用 FASTA3 程序进行相似性搜索

自从 10 年前发表以来[7]，FASTA 程序软件包有了显著的改进。原来的软件包提供了 4 个程序：fasta、tfasta、lfasta 和 rdf (参考文献[8]介绍了 rdf 与第一个 fastp 程序)。现在已经有了进行严格的 Smith-Waterman 搜索的程序(ssearch3)还有为混合的多肽序列的搜索程序(fastf3 和 tfastf3)；为翻译 DNA：蛋白质序列并比较的程序随着 fastx3、fasty3、tfastx3 和 tfasty3 的出现而有了实质上的改进，还有估计换位序列相似性得分统计学意义的程序(prss3)，它能进行精确的统计学分析。寻找的数据库的 FASTA3 程序总结在表 10.1，计算统计学意义的程序见表 10.2。

表 10.1 FASTA3 程序包中的比较程序

fasta3	把蛋白质序列和一个蛋白质序列数据库比较或 DNA 序列和一个 DNA 序列数据库比较使用 FASTA 算法[4,7]。搜索速度和选择以 *ktup* 参数控制(word 大小)。蛋白质比较时，缺省的 *ktup* = 2；*ktup* = 1 更敏感但是更慢。DNA 比较时，缺省的 *ktup* = 6；*ktup* = 3 或 *ktup* = 4 提供更高的敏感性；*ktup* = 1 应该被用于寡核苷酸(DNA 查询长度<20)
ssearch3	用 Smith-Waterman[22]算法把一个蛋白质序列和一个蛋白质序列数据库比较。ssearch3 的速度约是 FASTA3 的 1/10，但是在完整长度蛋白质序列比较时更敏感
fastx3/fasty3	把 DNA 序列和一个蛋白质序列数据库比较，通过比较在 3 个框架中的经翻译的 DNA 序列，并且允许间隙和框移。fastx3 使用更简单，更快的算法进行仅仅密码子之间移框的比对；fasty3 更慢，但是能产生差质量序列间好一些的比对，因为允许密码子中的移框
tfastx3/tfasty3	把蛋白质序列和一个 DNA 序列数据库比较，计算有正反方向框移的相似性
tfasta3	把蛋白质序列和一个 DNA 序列数据库比较，计算和 3 个正向以及 3 个反向读框的相似性(没有框移)。因为它们能计算移框时的相似性，tfastx3 和 tfasty3 常采用
fastf3	把混合的肽序列和一个蛋白质序列数据库比较。肽的混合物，典型地在氰溴化物切开以后未经进一步的分离由 Edman 降解获得，与一个数据库蛋白质序列作比较以鉴定那些很可能生产肽混合物的序列
tfastf3	把混合的肽序列和一个翻译的 DNA 序列数据库比较

表 10.2 FASTA3 程序包中的统计程序

prss3	由比较 2 条序列并且计算最佳的相似性分数以评价蛋白质或 DNA 序列相似性分数的显著性，然后重复打乱第 2 条序列，并且使用 Smith-Waterman 算法计算最佳的相似性分数。极值分布的特征参数从打乱的序列分数中估计并且用于计算未打乱序列相似性分数的统计学意义
sc_to_e	从未加工的分数、序列长度、从搜索过程估计出的统计学参数以及数据库的大小计算一个相似性分数的统计学意义
randseq	产生与查询序列一样长度和氨基酸组成的随机序列。随机序列在计算统计估计精确性中很有用。总的来说在数据库搜索时，匹配到一条随机查询序列的最高值期望值 *E* 应该约为 1

另外，在 FASTA2 软件包的若干程序还没包括到 FASTA3 程序中(表 10.3)。当作者写本章时(1998 年夏)，lalign 不是在 FASTA3 软件包中的 FASTA2 软件包中最重要的程序。lalign(和相关的图形程序 plalign 和 flalign)能把两个相似的蛋白质序列进行多重局部比对，而 FASTA3 和 FASTA 仅能进行单序列比对。多重局部比对能显示蛋白质的域，即蛋白质可能包含与库中序列共有强相似性的若干域。当有多重的相似性域时，FASTA3 仅显示出最相似的比对结果；lalign 则能检测其他的比对。

总的来说，如果 FASTA3 有你需要的功能，FASTA3 软件包的程序则比更旧的 FASTA2 程序好。FASTA3 软件包的程序有更具功能的统计学估计和错误处理，更多的积分模型(FASTA3 除了有 FASTA2 中的 PAM250、BLOSUM50 外，还有 BLOSUM62、MDM10、MDM20、PAM120 和 BLOSUM80)，还有一个更宽的比较函数阵列(fasty3、fastf3、tfasty3 和 tfastf3)。

表 10.3　仅在 FASTA2 中有的程序

程序	说明
lalign/plalign/flalign	用 Waterman-Eggert[24]算法的 sim 方程[23]在 2 条蛋白质或 DNA 序列之间找出多个局部比对。lalign 显示出传统型的比对；plalign 生产图形，而 flalign 生产 GCG 图程序的图形命令。这个程序进行连续的全 Smith-Waterman 比对，并且最好用于蛋白质比对。DNA 比对可使用 lfasta(见下)
lfasta/plfasta/flfasta	使用 fasta 算法在 2 条蛋白质或 DNA 序列之间找出多个局部比对。lalign 与一 flfasta 使用带局部条带比对的试探性(heuristic)fasta 算法。lalign 较常用于蛋白质比对，但是 lfasta 对很长的 DNA 序列来说快得多。plfasta 和 flfasta 产生图形输出
prdf	像 prss3，但是使用 fasta 算法而不是 Smith-Waterman 算法。prss3 比较受欢迎
align	采用线性的空格[25]全局比对 2 条蛋白质或 DNA 序列
aacomp	报告蛋白质序列的氨基酸组成和分子质量
grease/tgrease	使用 Kyte-Doolittle 方法计算蛋白质序列的亲水图[26]。tgrease 生产 tektronix 图形

10.2.1　程序的选择

使用 FASTA 程序进行蛋白质和 DNA 数据库搜索的许多研究者不熟悉程序包中的其他程序，或不清楚这些程序在什么时候应该使用。表 10.4 提供了使用在 FASTA3 程序包中程序的一些策略。

表 10.4　程序选择方法

问题	程序	解释	可选
鉴定未知蛋白质	(1) fasta3	一般蛋白质比较。要求速度时 *ktup* = 2(缺省)；要求更敏感的搜索时 *ktup* = 1；宜先在可能得到最小的库中搜索(如 SwissProt 而非 Genpept)	blastp
	(2) ssearch3	是 fasta3 的 1/10 至 1/50，但是提供最大的敏感性，DNA 比较时无优势	fasta3/blastp
	(3) tfastx3/tfasty3	假如在蛋白质库中找不到同源物，用 tfastx3 或 tfasty3 在 DNA 库中查找。tfasty3 提供更准确的比对，但是约慢 33%	tblstn/tfasta[a]
鉴定结构 DNA 序列	fasta3	如果 DNA 序列编码一蛋白质，则先用蛋白质序列比较，然后试以翻译的蛋白质序列比较(fastx3/fasty3)。重复 DNA 序列或结构 RNAs，则先以 *ktup* = 6(缺省值)搜索，然后 *ktup* = 3。*ktup*<3 仅在很短的序列时使用(PCR 引物)	blastn
鉴定 EST 序列	fastx3/fasty3	蛋白质序列比较远比 DNA 序列比较敏感，所以先看看是否 EST 编码一个和已知蛋白质同源的产物	fasta3/blastx/tblastx
鉴定新的序列	tfastx3/tfasty3	如可能，从同种的库中搜索 EST 序列。采用低/紧密 MDM20 记分矩阵检测紧密相关性并避免远相关性	tblastn/tblastx
确认统计学意义	prss3	采用 500~2000 次顺序打乱，必须把统计学意义以原先搜索过的数据库大小进行规范	
确认统计学估计	randseq	用于产生随机序列；然后用 fasta3(或 blastp 或 ssearch3)搜索，并使期望值 *E* 接近 1	

a. 不再推荐使用

表 10.4 的建议基于两条拇指规则(rules-of-thumb)：使用为你的问题所设计的程序；并且只要可能，在搜索 DNA 序列数据库前先搜索蛋白质序列数据库。蛋白质序列比较通常揭示 20 亿~30 亿年以前分叉的同源性序列；而对于 DNA 序列比较来说，回溯 2 亿~5 亿年都很困难。这样，蛋白质序列比较或翻译的 DNA 序列比较，允许人们鉴定在演化的时间上回溯 5~10 倍远时分叉的同源物(表 10.5)。

表 10.5　DNA 对蛋白质序列比较

最好分数		DNA *E*(188 018)	tfastx3 *E*(187 524)	蛋白质 *E*(331 956)
DMGST	*D. melanogaster* GST1-1	1.3e–164	4.1e–109	1.0e–109
MDGST1	*M. domestica GST-1* 基因	2e–77	3.0e–95	1.9e–76
LUCGLTR	*Lucilia cuprina* GST	1.5e–72	5.2e–91	3.3e–73
MDGST2A	*M. domesticus GST-2* mRNA	9.3e–53	1.4e–77	1.6e–62
MDNF1	*M. domestica nf1* 基因 10	4.6e–51	2.8e–77	2.2e–62
MDNF6	*M. domestica nf6* 基因 10	2.8e–51	4.2e–77	3.1e–62
MDNF7	*M. domestica nf7* 基因 10	6.1e–47	9.2e–77	6.7e–62
AGGST15	*A. gambiae GST* mRNA	3.1e–58	4.2e–76	4.3e–61
CVU87958	*Culicoides* GST	1.8e–41	4.0e–73	3.6e–58
AGG3GST11	*A. gambiae GST1-1* mRNA	1.5e–46	2.8e–55	1.1e–43
BMO6502	*Bombyx mori GST* mRNA	1.1e–23	8.8e–50	5.7e–40
AGSUGST12	*A. gambiae GST1-1* 基因	2.3e–16	4.5e–46	5.1e–37
MOTGLUSTRA	*Manduca sexta* GST	5.7e–07	2.5e–30	8.0e–25
RLGSTARGN	*R. legominosarum gstA* 和 *gstR*	0.0029	3.2e–13	1.4e–10
HUMGSTT2A	*H. sapiens* GSTT2	0.32	3.3e–10	2.0e–09
HSGSTT1	*H. sapiens GSTT1* mRNA	7.2	8.4e–13	3.6e–10
ECAE000319	*E. coli* hypothet 蛋白质	–	4.7e–10	1.1e–09
MYMDCMA	*Methylophilus dichlorometh.* DH	–	1.1e–09	6.9e–07
BCU19883	*Burkholderia maleylacetate* red.	–	1.2e–09	1.1e–08
NFU43126	*Naegleria fowleri* GST	–	3.2e–07	0.0056
SP505GST	*Sphingomonas paucim*	–	1.8e–06	0.0002
EN1838	*H.sapiens* maleylacetoacetate iso.	–	2.1e–06	5.9e–06
HSU86529	Human GSTZ1	–	3.0e–06	8.0e–06
SYCCPNC	Synechocystis GST	–	1.2e–05	9.5e–06
HSEF1GMR	*H. sapiens EF1*g mRNA	–	9.0e–05	0.00065

灵长类、其他哺乳动物、无脊椎类和细菌部分的 GenBank 用果蝇谷胱甘肽转移酶 cDNA(DMGST)和蛋白质(gtt1_drome)序列以 fasta3(DNA，*ktup* = 4)、tfastx3 和 fasta3(蛋白质，*ktup* = 2)搜索。所选择的高分值序列的期望值显示于表中。有“–”的 DNA 比较期望值 $E > 100$。这类查询中，DNA 序列比较仅仅在其他昆虫中检测同源物，而蛋白质和翻译的 DNA 比较发现与从人和细菌中同源物有统计学意义的相似性

另外，低复杂性区域相对容易从蛋白质序列数据库中移开，并且在蛋白质序列比对时容易被识别，但是要在 DNA 序列比对时识别出则要困难得多。由于其异常的氨基酸组成，这些区域能在非同源性序列间产生统计学上有意义的相似性分数。这样，当需要鉴定一新测序的表达序列标签(EST)序列时，你应该首先使用 fastx3 或 fasty3 搜索像 SwissProt 或 PIR 一样的较全面的蛋白质数据库，然后搜索像 BLAST/NCBI(生物技术信息国家中心)、nr 或 OWL[9]一样的更大但是更冗余的数据库，非冗余蛋白质数据库，或 Genpept，然后，只有在这些搜索以后没能获得统计上有意义的匹配时，才应该寻找 DNA 序列匹配。

10.2.2 FASTA 对 BLAST

序列比较程序的 BLAST 家族[10, 11]提供许多和 FASTA 程序一样的搜索能力(表 10.6)。总的来说，BLAST 程序更快，但是 FASTA 程序能进行更精确的比对。大多数蛋白质序列数据库搜索时，当前的 blastp2.0 (带间隙的 BLAST，参考文献[11])能像 fasta3 和甚至更严格的 ssearch3 一样有效地鉴定未知的蛋白质。从 blastp2.0 (BLOSUM62，在间隙中的第一残基 -12 分，其余每一残基 -1 分)起，fasta3 和 ssearch3 使用不同的记分矩阵(BLOSUM50)和间隙罚分(在间隙中的第一残基-12 分，其余每一残基 -2 分)。先前的 blastp1.4 生产很差的序列比对结果(因为在间隙上的限制)，但是 blastp2.0 版本进行与那些严格的 Smith-Waterman 搜索很相似的蛋白质序列比对。

表 10.6 比较 BLAST2 和 FASTA3 程序

BLAST	FASTA	功能
blastp	fasta3	蛋白质序列相似性搜索。blastp 更快，并且能在一样的序列中的若干域之间显示比对。fasta3 显示一 Smith-Waterman 的最后比对并且在一些情况下产生更精确的统计估计
blastn	fasta3	DNA 序列比较。blastn 的速度被高度优化；它使用固定的词大小(11 核苷酸)和对一些问题不适合的积分矩阵(例如，找 PCR 引物匹配时)
blastx	fastx3/ fasty3	把一翻译的 DNA 和一个蛋白质序列数据库进行比较。而 blastx 能进行 6 个独立的搜索(为 6 个框架中的每一个各进行一次搜索)，fastx3 和 fasty3 有效地进行单个正向(或反向)搜索在计算相似性分数时允许框移和比对。结果，fastx3 和 fasty3 更敏感并且当 DNA 序列有移框错误时，能产生比 blastx 好一些的比对
tblastn	tfastx3/ tfasty3/ tfasta	把蛋白质序列和一个 DNA 序列数据库比较，进行 3 个正向和反向框架的翻译。再者，当 DNA 序列有移框错误时，tfasty3 和 tfasty3 比 tblastn 或 tfasta 提供更精确的比对
	tblastx	把 DNA 查询序列和一个 DNA 文库比较，在所有的 6 个框架中翻译两序列并且使用一个蛋白质代矩阵(BLOSUM62)记分。*ktup* = 6 的 fasta3(缺省)提供相似的功能，但是不使用记分矩阵的蛋白质

对翻译推得的 DNA-蛋白质比较和 DNA 数据库搜索，FASTA 程序比 BLAST 对应程序好得多。尽管带间隙的 blastp2.0 在蛋白质比较中表现很好，blastx 独立

提供 3 个正向框架搜索，而 fastx3 和 fasty3 计算允许移框的单个比对。把所有 3 个正向读框当成单个的序列使其更易产生越过匹配的蛋白质序列长度的高质量的比对，并且允许从不同读框的相似性以自然的方法被延续以改进敏感性。例如，blastx 搜索小鼠μ-谷胱甘肽转移酶 cDNA 序列，当在 5%的位置有插入和删除错误时仅检测到其他的μ-谷胱甘肽转移酶，而同样序列使用 fasty3 搜索却检测到更多的μ类蛋白质序列[$10^{-20} < E() < 10^{-17}$]，还有其他 8 个关系更远的π-谷胱甘肽转移酶序列($10^{-5} < E < 0.01$)。

FASTA 程序也为 DNA 序列搜索提供附加的灵活性。搜索过程能采用任何的词大小(*ktup*)，从 1~6；小的 *ktup* 特别适于短序列的搜索，如 PCR 引物。另外 FASTA 程序能使用许多记分矩阵，包括带高错配罚分的矩阵，用于在序列中鉴定长的特征序列。

10.3 解释 FASTA 统计

在 1983 年快速序列比较程序第一次出现后[1]，经过寻找序列数据库发现相似的 DNA 和蛋白质序列变得可能，但是弱相似性是否可能有生物学意义的判断还没有正式的原则。FASTP 程序提供了为计算相似性得分的 Monte-Carlo 打乱法(rdf)[8]，但是为有意义的相似性($Z > 5$)，推荐采用并不基于局部的相似性分数的正确统计模型，且不考虑数据库大小。一个得分高于平均值 10 个标准误差($Z > 10$)的序列，于一个 10 000 条数据库中被搜索到的概率是 0.015，SwissProt 中(70 000 条数据)为 0.11；这样在统计上无意义，甚至在 0.05 的水平上。

blastp 程序的相似性搜索集成了精确的统计学估计[10]，基于局部的相似性分数能精确地通过极值分布描述达到[12,13]。fastp 中的 Monte-Carlo 打乱法程序使用极值分布计算比对分数概率，并且在 FASTA2 和 FASTA3 程序包的文库搜索程序提供能被用来从有统计意义相似性的值期望(*E*)值推断同源性[6]。

E 值是决定是否进一步分析一高等级序列比对时，你应该看的第一个数字。研究者经常想知道他们应该使用什么 *E* 值。这将在下一节中详细讨论，但是在大多数情况下，并且在 0.001~0.01 之间的 *E* 值能用来可靠地推断同源性，但要求进行成百上千个搜索时，则需要更低(更保守)的值(如鉴定一个细菌染色体的所有基因)。

由 FASTA3 程序和 BLAST 程序计算的 *E* 值是所观察的相似性分数随机发生的可能性的一项统计学量。像任何统计学量一样，它的实用性取决于：内在的统计模型的假设是否正确，以及当使用量值得出一个结论时，可以接受的错误类型。进行相似性搜索时，我们从统计上有意义的相似性推断同源性(共同的祖先)。然而，统计学意义的阈值会变化，取决于我们是否更关心偶发的错误识别一个同源物(当一个非相关的序列标记为相关，假阳性或I类误差)或错过一可能的同源物(当

发现一高分值同源物时，却把序列标记为非同源性，假阴性或类型 II 错误)。

10.3.1 推断序列同源性的阈值确定方法

对于大多数分子生物学家来说，相似性搜索中最关心的是假阳性错误，我们不想送一封信到 Nature 发表有关鉴定出 p53_human 的酵母同源物，却没有进化过程的亲缘存在。(同源性测试的金标准是结构的相似性。如果 p53 的候选酵母同源物有完全不同的三维结构，那么假设就是错误的。)在精确的相似性统计可用之前不正确的同源性判定较常见，现在却很少见。(然而很不幸，一旦结果出版了，从出版物中去除就困难了) E 值或 fasta3 计算的期望值是在搜索数据库中期望看到一个分数时的次数等于或大于此值。换句话说，$E<0.01$ 即你期望看到在 100 次搜索中看到某分数一次(或更多)；$E<0.001$ 即在 1000 次搜索中有一次，如此等等。E 约为 1，即期望每次搜索都能看见某一个分数。

blast 程序的更早版本使用了一相关统计 p 值，表示一个相似性分数的显著性。由 fasta 程序报告的 E 值范围从 0~D，D 在此是数据库的记录数，而 blast p 值范围是 0~1。E 值的概率[$p()$值]能用泊松分布公式计算：$p(E) = 1-e^{-E}$。$E<0.1$ 时的值，$p \sim E$，故 $p(E = 0.1) = 0.1$；$p(E = 1.0) = 0.63$；$p(E = 5.0) = 0.99$。

而有意义的 E 阈值(0.001~0.01)能保证研究人员避免假阳性错误。避免假阴性的方法很少，即把一个序列标记为与数据库中的任何记录无关，而事实上有一同源物存在。大多数多种多样的蛋白质家族包含相关序列对，而在序列相似性上无显著的统计学意义。幸好，如果那些家族很大(例如，球蛋白、丝氨酸蛋白酶、谷胱甘肽转移酶、G 蛋白耦联受体)，那么很有可能新发现的家族成员将与一些已知成员间有显著的相似性。当序列数据库变得更完全并且蛋白质家族变大后，假阴性率应该减少。

10.3.2 选择一个数据库

相似性分数的期望值 $E(S>x)$是从成对的相似性分数的概率 $p(S>x)$计算得出，它能使用极值分布[12,13]和发现高分数序列所进行测试的次数(即序列比较)进行计算。这样，$E(S>x) = p(S>x)D$，D 在此是在数据库中序列的数目(DNA 序列比较时，D 不是在数据库的序列的数目，但在数据库中核苷酸长度由查询序列的长度划分)。

因为 E 随着数据库记录数线性地增加，在搜索一个有 1000~5000 个记录的细菌染色体时发现的相似性将比在 OWL 非冗余蛋白质数据库中一个完全相同分数比对强 50~250 倍(参考文献[9]；250 000 条记录)。这样，当在找相隔很远的关系时，应该始终使用可能包含所感兴趣的同源物的最小数据库。如果目标是寻找大肠杆菌巴氏杆菌 DAHP 合成酶(arog_bacsu)的同源物，应该搜索大肠杆菌蛋白质组(数据库)[可找到大肠杆菌 kdsA 同系物且 $E(4283)<0.00015$]而不是 SwissProt

[kdsa_ecoli E(74 417)<0.0017]或 OWL[kdsa_ecoli E(260 784)<0.0085]。这里，采用同样的比对与相似性记分，其显著性在大数据库中是最小数据库的 1/50。

同样，搜索 SwissProt (约 70 000 条记录)将比搜索 OWL (261 000 序列)或 BLAST nr 蛋白质数据库(332 000 序列)敏感性强 3~5 倍，仅因为 SwissProt 更小。因此，鉴定蛋白质同源物的有效策略应该是: ①首先寻找更小的数据库；②然后用更敏感的算法($ktup = 1$ 时的 fasta3 或 ssearch3)研究一个更小的数据库(像 SwissProt)，如果没发现有意义的匹配，则③搜索更大的数据库(OWL 或 nr)。然而数据库的大小会减小搜索灵敏度，数据库提供某蛋白质家族更多样的成员时，更大的数据库可能有效。例如，在 SwissProt 库中，最远缘的 `p53_human` 同源物是一个较难搜索到的序列。而 OWL 包含 2 倍多的新 p53 同源物，包括鱿鱼的一个同源物。

10.3.3 大规模序列分析的阈值

染色体测序中心和每天做几千个相似性搜索的其他小组必须使用有统计学显著性的更保守的阈值以避免假阳性错误。$E = 0.001$ 的阈值，对于每天做数个搜索的某些人来说有点保守，应该在 10 000 搜索以后，在非同源性序列之间产生低于阈值的 10 个分数。确实，如果对 PIR 或 SwissProt 数据库以随机序列做 100 次搜索，这 100 个序列其中之一将在 $E<0.01$ 时找到一条同源物，$E<0.1$ 时会找到 10 条，如此等等[6]。染色体测序中心典型地使用 $E<10^{-6}$，或更低的值作为鉴别数千条序列的阈值。

然而，当查找远缘关系时，使用更保守的具有统计显著性的阈会使你得到更多假阴性(II 类)错误。例如，在用从 SwissProt 中来的 2608 条人的蛋白质进行对大肠杆菌蛋白组(4289 序列)比较，417 条获得了 $E<0.02$，373 条 $E<0.01$，301 条 $E<0.001$，256 条 $E<0.0001$。在 72 条 $0.001<E<0.01$ 的序列中，我们可以期望大约 26 条(0.01 2608)具有这样高的相似性的概率较小，而其他 45 条是真正同源的。(不幸的是，没有附加信息时我们不能确认哪 45 条序列是同源物)在人/大肠杆菌搜索中，209 条序列 $E<10^{-6}$，我们可以期望所有这些匹配具备真正同源性。然而，采用保守的 10^{-6} 阈将会使近 200 条可能的同源物误鉴定成无关序列。这样，在新测序的细菌染色体中，对新的或未鉴别出的蛋白质的数量通常过高估计，因为许多新的蛋白质可能在个别查找时有显著的相似性，然而当在 2000~4000 条序列一组查找时却没有。

10.3.4 统计学估计——你能信任什么？

如果统计学的估计是精确的，在前几章的指南中就会为基于序列相似性鉴定相关序列提供可靠的策略。然而，对于生物学序列(与公平的硬币对比)，统计学模型中的假设一般达不到。当假设失败时，最高分数的无关序列可以有太低的，

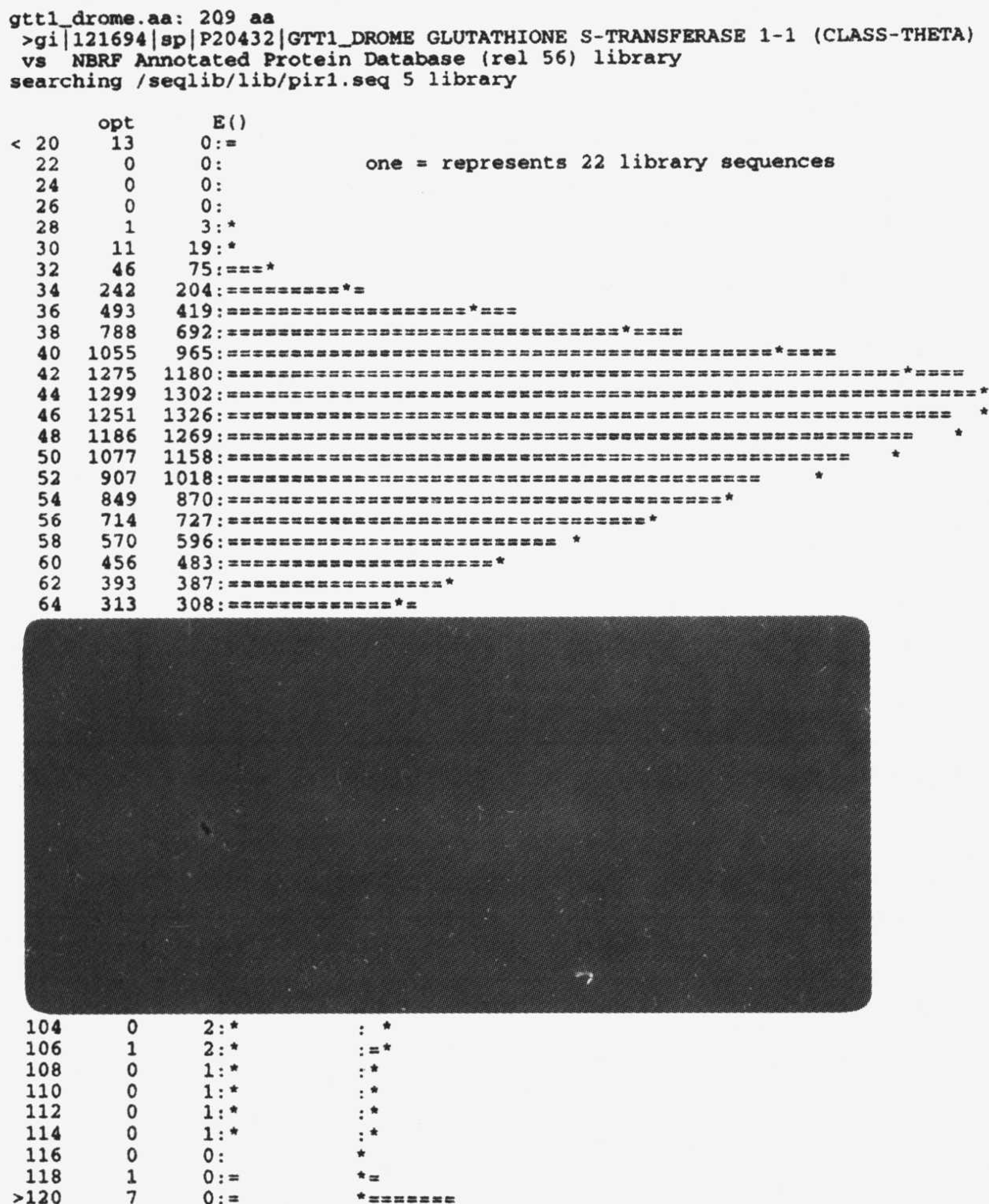
如[$E<10^{-3}$]或太高的[$E>100$]期望值。如果 E 值太低，无关的序列将错误地被标记为相关(假阳性)。如果 E 值太高，可能同时相关序列的 E 值也太高，这样相关序列将被错过(假阴性)。

总的来说，不准确的统计学的估计是由错误的间断罚分或查询序列中低复杂性区域引起(单一种类氨基酸组成的序列，例如，*Caenorhabditis elegans* 胶原的 ggqgppgdaggpg 序列或果蝇胰蛋白酶的 ssggvtfsvss)[3, 14]。在第一种情况下，统计学模型无效。在 BLASTP、FASTA 和 Smith-Waterman(ssearch3)的分数估计后面的统计学理论假设分数是局部的，即平均地、不一样的氨基酸的相似性分数 $s_{ij}<0$。如果间隙罚分太低，那么比对算法将选择插入间断，而非结束比对，并且比对将趋于全局性，即从头到尾进行比对。全局比对分数的统计学性质与局部比对分数不同。局部分数跟随极值分布；全局比对分数的分布尚待进一步探讨。

序列统计的可靠性可以通过直方图快速确认，在 FASTA3 搜索和显示的相似性分数的观察值及期望值的直方图，还可以通过核对高分值非相关序列的期望[E]值来确认。有些例子显示运行 FASTA3 和 ssearch3 程序的结果，可以在 ftp://ftp.virginia.edu/pub/fasta/站点上看到。这个站点上的程序可以在大多数的 UNIX 平台上运行(如数字公司的 UNIX、IBM、AIX、Linus、SGI Irix 以及 Sun Solaris 平台)，此外，还能在 Windows(Windows95/NT)和 Macintosh 平台上运行。这里显示的输出稍微与遗传学计算机组[Genetics Computer Group(GCG)]分发的 FASTA 程序不同，但是序列相似性信息在所有的新版本 FASTA 都一样。尽管计算最高分值非相关序列似乎让我们知道了蛋白质家族的特性，但是用其他候选的非相关序列[E~1]搜索常常能分离低分数相关序列和高分值无关序列[5]。如果在实际和期望分值分布之间有良好协议，并且最高分值非相关序列的 $E()$值约为 1 的话，统计学估计就会很精确。

10.3.4.1 低差距罚分引起不精确的估计

大多数蛋白质和 DNA 序列搜索时，在观察到的和期望的分数的分布(图 10.1)之间有很好的协议并且高分数的无关的序列 E 值约等于 1(表 10.7，参考文献[6])。FASTA 程序在每次搜索以后提供一张总结观察到的和期望的分数分布的直方图(图 10.1~图 10.3)。图 10.1 报告了一次搜索的结果，数据库中的 788 条序列(选择柱)获得了 38~39 分(最左边的柱形)，而 692 条序列[E 柱]被期望在一个 14 000 条序列的数据库中落在这个分数范围之间。在观察到的(“ === ” 图)和期望的(在直方图的“*”)之间的协议在图 10.1 的阴影区域是特别重要的。在许多搜索中，由高分值序列列表中查找高分值无关序列来确定估计的精确度也是可能的。在表 10.7 高分值无关序列是 S30223 和 NOBY2，期望值约等于 8。(理想状态下，此分值应该接近 1；用 ssearch3 搜索得到的高分无关序列 $E<3$。)



```
gtt1_drome.aa: 209 aa
 >gi|121694|sp|P20432|GTT1_DROME GLUTATHIONE S-TRANSFERASE 1-1 (CLASS-THETA)
 vs  NBRF Annotated Protein Database (rel 56) library
searching /seqlib/lib/pir1.seq 5 library

       opt      E()
< 20    13      0:=
  22     0      0:          one = represents 22 library sequences
  24     0      0:
  26     0      0:
  28     1      3:*
  30    11     19:*
  32    46     75:===*
  34   242    204:=========*=
  36   493    419:===================*===
  38   788    692:===============================*====
  40  1055    965:===========================================*====
  42  1275   1180:=====================================================*====
  44  1299   1302:===========================================================*
  46  1251   1326:=========================================================  *
  48  1186   1269:======================================================   *
  50  1077   1158:=================================================   *
  52   907   1018:==========================================    *
  54   849    870:=======================================*
  56   714    727:=================================*
  58   570    596:========================== *
  60   456    483:=====================*
  62   393    387:=================*
  64   313    308:=============*=
 104     0      2:*          : *
 106     1      2:*          :=*
 108     0      1:*          :*
 110     0      1:*          :*
 112     0      1:*          :*
 114     0      1:*          :*
 116     0      0:           *
 118     1      0:=          *=
>120     7      0:=          *=======
```

图 10.1　果蝇 theta 类谷胱甘肽转移酶(gtt1_drome)以 fasta3 程序在标注的 PIR1 蛋白质序列数据库中搜索的相似性分数结果直方图

显示起始的直方图输出，阴影部分表示当统计的模型失败时，最可能显示观察和期望分值数目差异的区域

表 10.8 和表 10.9 以及图 10.2 显示统计模型失败了的搜索的 2 个例子。在第一个例子中(表 10.8)，DNA 搜索时的间距罚分为−12 和−2，而非缺省的−16，−4。而直方图(未显示出)显示出在观察到的和期望的分数的分布之间有很好的协议，高分数的无关的序列 *E* 值是 0.01。(tfasty3 扫描证实高分数无关的序列不包含一同源物。)而且，当间距罚分从−16/−4 减小为−12/−2 后，同源性比对的 *E* 值增加 10^7(例如，AC002520 从 1.2×10^{-12}~0.0008，表 10.8)。有甚至更低的间距罚分的 DNA 序列搜索确实显示出在观察到的和期望的分数的分布之间可观的差别，但是高分数的无关序列的 *E* 值通常是和统计学估计精确性有关的最灵敏的度量。

表 10.7 FASTA 搜索——高分值序列

名称	描述	长度	initn	opt	z–分值	E()
XUFF11	glutathione transferase	209	1399	1399	1626.5	1.2e–84
XUZM32	glutathione transferase	222	133	173	210.9	8.6e–06
XUZM31	glutathione transferase	220	107	164	200.6	3.2e–05
XUZM1	glutathione transferase	213	123	144	177.7	0.00061
RGECSS	string. starv. prot.-*E. coli*	212	106	140	173.1	0.0011
XURTG	glutathione transferase	222	58	139	171.7	0.0013
XURT8C	glutathione transferase	222	39	115	144.0	0.046
XURTG4	glutathione transferase	218	40	93	118.7	1.2
A37378	glutathione transferase	210	40	82	106.2	5.8
S30223	Elongation factor eEF-1g	227	34	80	103.5	8.3
NOBY2	*phosphopyruvate hydratase*	437	53	83	103.1	8.8
PWBYD	*H+-transporting ATP synthase*	212	53	79	102.7	9.2

fasta3 在标注的 PIR1 数据库[27]中搜索 `gttl_drome` 得到的高分值序列(*ktup* = 2)，高分值的不相关序列以斜体表示

表 10.8 FASTA 搜索——低间隙罚分

最好的分值是:		长度	initn	opt	z–分值	E(−12/−2)	E(−16/−4)
AC002520	Human Chr. 1p13	11 901	1507	404	173.1	0.0008	1.2e–12
AC000031	Human Chr. 1p13.3	39 043	1396	394	161.0	0.0011	6.5e–12
HSU47924	*Human chr. 12p13*	78 864	235	352	138.3	0.01	2.0
AC000032	Human Chr. 1p13	29 867	1354	345	141.6	0.018	6.6e–09
CACD42	*C.atys CD4 mRNA*	1189	69	307	146.1	0.26	—
HUMDXS455A	*Human cosmid*	38 409	126	274	109.2	0.89	—
HSHS12ENH	*Homo sapiens DNA*	3735	151	278	126.1	1.1	0.038
HSV411C11	*Human DNA*	5637	165	276	122.5	1.1	—
HUMHSLA	*Human hormone-sens.*	3255	63	275	125.7	1.3	—
AF031078	*Human chr. X*	78 864	188	264	100.2	1.4	0.078
AF035180	*Human chr. 4q35*	4638	67	271	121.7	1.5	0.08

fasta3 搜索(*ktup* = 6)GenBank106(约 80 000 条序列)得到的高分值序列，所采用的是 mGstm1 cDNA 序列(MUSGLUTA)的反向互补形式，缺省的替换矩阵(+ 5/−4)和低(−12/−2)或缺省的(−16/−4)间隙罚分。不相关序列用斜体加亮显示。低间隙罚分改善了不相关HSU47924序列的*E*值为*E*<0.01，并且减小了同源的AC002520、AC000031和 AC000032 序列的同源性 10^{7}

表 10.9 FASTA 搜索——低复杂度区域

以完整的 grou_drome 搜索:		长度	initn	init1	opt	z-分值	E(14 212)
RGHUB1	GTP-binding reg. prot.	340	161	147	237	197.4	4.9e–05
RGHUB3	GTP-binding reg. prot.	340	163	152	233	194.2	7.4e–05
RGBOB2	GTP-binding reg. prot.	326	181	149	228	190.5	0.00012
PIHUB6	*salivary proline-rich prot*	392	142	142	229	190.1	0.00013
RGKWB	GTP-binding reg. prot.	340	159	154	222	185.4	0.00023
RGFFBH	GTP-binding reg. prot.	340	169	144	219	183.0	0.00031
PIHUSD	*proline-rich glycoprot.*	310	141	141	217	182.0	0.00035
PIRT3	*acidic proline-rich protein*	206	138	138	212	180.7	0.00042
WMBEW6	*capsid protein-herpes*	635	101	101	206	168.7	0.002
S23447	*annexin XI form B-bovine*	505	84	84	202	166.9	0.0024
PIHUPF	*salproline-rich glycoprot.*	251	147	147	193	164.3	0.0034
PIHUSC	*proline-rich phosphoprot.*	166	88	88	180	156.6	0.0092
CGHU6C	*collagen alpha1(II)*	1487	104	104	197	156.0	0.0099
RGOOBE	GTP-binding reg. prot.	341	156	125	181	152.8	0.015
FOLJSP	*gag polyprotein-foamy vir*	811	121	121	187	151.9	0.017
CGBO1S	*collagen alpha 1(I) -bovine*	779	88	88	185	150.6	0.02
LUDO7	*annexin VII-slime mold*	462	88	88	179	149.2	0.024
CGHU2S	*collagen alpha 2(I)*	1366	88	88	187	148.6	0.026
LUBO11	*annexin XI form A-bovine*	503	84	84	177	147.1	0.031
S09257	*Hox A4-chicken*	309	116	116	172	146.2	0.035
OZZQMY	*circumsporozoite prot pre.*	367	146	146	172	145.1	0.04
以 grou_drome 的部分序列搜索:(低复杂度区域已移去)							
RGHUB1	GTP-binding reg. prot.	340	161	147	237	247.5	8e–08
RGHUB3	GTP-binding reg. prot.	340	163	152	233	243.3	1.4e–07
RGHUB2	GTP-binding reg. prot.	340	181	149	228	238.1	2.7e–07
RGKWB	GTP-binding reg. prot.	340	159	154	222	231.9	5.9e–07
RGFFBH	GTP-binding reg. prot.	340	169	144	219	228.7	8.9e–07
RGOOBE	GTP-binding reg. prot.	341	156	125	181	189.1	0.00014
BVBYMS	*MSI1 protein-yeast*	422	116	74	139	143.9	0.047
ERHUAH	*coatomer complex alpha*	1224	109	109	134	131.7	0.23
I37062	*involucrin S-gorilla*	495	129	81	115	117.8	1.3

不相关序列以斜体标亮

```
grou_drome.aa: 719 aa
 >GROU_DROME GROUCHO PROTEIN (ENHANCER OF SPLIT M9/10). - DROSOPHILA MELANOGAS
 vs  NBRF Annotated Protein Database (rel 56) library
searching /seqlib/lib/pir1.seq 5 library

      opt      E()
< 20   13     0:=
  22    0     0:          one = represents 28 library sequences
  24    1     0:=
  26    0     0:
  28    1     3:*
  30   10    20:*
  32   21    76:= *
  34  105   205:====   *
  36  272   422:==========     *
  38  540   697:====================    *
  40  937   972:==================================*
  42 1269  1188:==========================================*===
  44 1645  1311:==============================================*============
  46 1666  1335:===============================================*============
  48 1577  1278:=============================================*===========
  50 1310  1166:=========================================*=====
  52 1056  1025:====================================*=
  54  851   876:===============================*
  56  669   732:========================  *
  58  423   601:================     *
  60  419   487:===============  *
  62  255   390:==========   *
  64  196   310:=======    *
  66  181   245:======= *
  68  154   193:======*
  70   99   151:==== *
  72   74   118:=== *
  74   63    92:===*
  76   60    72:==*
  78   47    56:=*
  80   48    43:=*
  82   36    33:=*
  84   33    26:*=
  86   27    20:*
  88   21    16:*          inset = represents 2 library sequences
  90   18    12:*
  92   20     9:*         :====*=====
  94   20     7:*         :===*======
  96   17     6:*         :==*======
  98    7     4:*         :=*==
 100   10     3:*         :=*===
 102   11     3:*         :=*====
 104   10     2:*         :*====
 106   11     2:*         :*=====
 108    7     1:*         :*===
 110   10     1:*         :*====
 112    6     1:*         :*==
 114    4     1:*         :*=
 116   11     0:=         *======
 118   10     0:=         *=====
>120   70     0:===       *===================================
5446221 residues in 14321 sequences
 Expectation_n fit: rho(ln(x))= 8.0964+/-0.00108; mu= 4.7475+/- 0.061;
 mean_var=157.6967+/-31.622, 0's: 13 Z-trim: 96  B-trim: 33 in 1/62
 Kolmogorov-Smirnov  statistic: 0.0497 (N=29) at  52
```

图 10.2 低质量统计:低复杂性区域——fasta3 采用 grou_drome 搜索(*ktup*=2)PIR1 数据库。显示出序列相似性分数的直方图。在这种情况下，观察和期望的序列数有显著差异，这些序列的分值在分布的中间部分和尾部，高分值序列过多。表 10.9 显示这些高分值序列是非相关的

10.3.4.2 来源于低复杂度区域的低 *E*()值

在非同源性序列之间的低 *E*()值通常由低复杂性区域引起[3,14]。果蝇 groucho 蛋白质序列(grou_drome)仅包含 5 个低复杂性区域(seg 程序从 719 个残基中鉴定出 83 个，参考文献[14])，但是通过比较图 10.2 和图 10.3 显示，在这些区域的配对明显地扭曲了高分值无关序列的分布。相反，隐蔽了 5 个低复杂度区域的搜索(图 10.3)显示出所期望的分值分布。检查低复杂性搜索中高分值序列表(表 10.9)可以看出大量的和无关蛋白质有意义匹配[$0.00013<E<0.02$]，这些蛋白质都带有氨基酸

组成偏性，然而用 seg 处理过的搜索中最高分值的无关序列的 E 值小于 0.047。令人惊奇的是，相关 GTP-结合调节蛋白相似性分值的显著性提高了几乎 1000 倍(表 10.9)。

```
grou_drome. Seg: 719 aa
 >GROU_DROME GROUCHO PROTEIN (ENHANCER OF SPLIT M9/10). - DROSOPHILA MELANOGAS
 vs  NBRF Annotated Protein Database (rel 56) library
searching /seqlib/lib/pirl.seq 5 library

       opt      E()
< 20    48      0:==
  22    14      0:=          one = represents 24 library sequences
  24    21      0:=
  26    37      0:==
  28    39      3:*=
  30    65     20:*==
  32    95     76:===*
  34   175    206:========*
  36   348    424:===============  *
  38   591    700:=========================    *
  40   891    977:======================================  *
  42  1141   1194:================================================ *
  44  1328   1317:======================================================*=
  46  1373   1342:=======================================================*==
  48  1395   1285:=====================================================*=====
  50  1227   1172:================================================*===
  52  1107   1031:==========================================*====
  54   888    880:====================================*
  56   723    735:==============================*
  58   602    604:=========================*
  60   490    489:====================*
  62   357    392:=============== *
  64   284    312:============*
  66   246    246:==========*
  68   177    194:========*
  70   131    152:======*
  72   110    119:====*
  74    64     93:===*
  76    76     72:==*=
  78    53     56:==*
  80    41     43:=*
  82    44     33:=*
  84    22     26:=*
  86    26     20:*=
  88    17     16:*            inset = represents 1 library sequences
  90    11     12:*
  92    14      9:*         :========*=====
  94     5      7:*         :===== *
  96     7      6:*         :=====*=
  98    11      4:*         :===*=======
 100     2      3:*         :==*
 102     5      3:*         :==*==
 104     3      2:*         :=*=
 106     1      2:*         :=*
 108     1      1:*         :*
 110     0      1:*         :*
 112     1      1:*         :*
 114     0      1:*         :*
 116     0      0:           *
 118     1      0:=          *=
>120    13      0:=          *=============
5446221 residues in 14321 sequences
 Expectation_n fit: rho(ln(x))= 6.3481+/-0.00105; mu= 10.5411+/- 0.059;
 mean_var=92.0111+/-17.844, 0's: 13 Z-trim: 24  B-trim: 593 in 1/62
                   Kolmogorov-Smirnov  statistic: 0.0129 (N=29) at  42
```

图 10.3　用“seg-ed”查询的精确统计——在图 10.3 的搜索中采用的是掩盖了低复杂度的 grou_drome 序列(用 seg 程序修饰过)[14]。由于低复杂性序列去除了，观察到的和所期望的相似性分数的数目很吻合。当低复杂度区域从 PIR1 数据库中去除后得到了相同的结果

在蛋白质-蛋白质数据库搜索中，去除低复杂度序列对于查询序列或数据库序列是等效的。但是，从 DNA 查询序列中去除低复杂度区域更困难，如 EST 序列。不幸的是，低复杂度蛋白质序列和读框外 DNA 翻译序列之间的比对是很常见的 [15]。改善用 DNA 翻译产物搜索的一个简单策略就是采用 seg 程序处理过的蛋

白质数据库[14]。

低间隙罚分和低复杂度区域产生不可靠的统计学估计，其原因就是隐含的统计学模型的假设不适宜。低间隙罚分引起比对从局部漂移到全局，极值比对统计仅适用于局部比对。低复杂度区域违反了有关无关序列中更高级别结构所隐含的假设。在低复杂度序列中有统计学意义的匹配没有生物学意义，因为统计学模型认为随机(无关)序列的每一位点和其他位点没有联系。

当统计学模型有效时——局部比对和真正随机无关序列存在，此时统计学意义显著的相似性分值能可靠地用于推断同源性。也可以经常核对统计学模型是否正确，这是通过观察实际的和期望的相似性分值的直方图，以及检查最高分值的无关序列的期望值来判断。

10.4 FASTA3 程序选项

在 FASTA 软件包中的程序行为能用许多命令行选项来修改；这些选择包括改变计分矩阵和间隙罚分、使用其他统计学估计算法和改变比对输出的格式。许多选项适用于软件包中的所有程序(表 10.10)；另外的选项仅用于 fasta3 或 tfastx/y3(表 10.11)。

表 10.10 FASTA3 一般选项

-a	显示所有序列，而不是仅仅重叠区域(fastx/y3 和 tfastx/y3 不提供此功能)
-b #	显示的最佳分值数目(必须小于-*E* 的取值点)
-d #	显示的最佳比对数目(必须小于-*E* 的取值点)
-E #	显示分值和比对的期望值边界。(缺省值为：蛋白质序列比较为 10.0，fastx/y3 为 5.0，DNA 为 2.0)
-H	关闭直方图显示
-i	(仅 DNA 查询)用查询序列的反向互补形式搜索(tfastx/y3)；仅和库序列的反向互补形式比较
-L	在比对中报告长序列描述
-m 1~6，10	比对显示选项(表 10.14)
-n	使程序查询核苷酸序列(缺省：自动检测)
-N #	以#个残基的块读取数据库。#必须大于查询序列的 2 倍，因为数据块的重叠长度等于查询序列的长度。(缺省值：80 000 查询长度)
-O file	输出结果到文件
-q/-Q	沉默。即不提示输入
-R file	把所有输入保存到统计文件中
-S #	矩阵值的偏移替换
-s name	记分矩阵。BLOSUM50 是蛋白质的缺省算法，PAM120、PAM250 和 BLOSUM62 可以由设置-s P120、P250 或 BL62 指定。其他矩阵包括：BLOSUM80(BL80)和 MDM_10、MDM_20、MDM_40(M10、M20、M40，参考文献[19])。此外还可以指定 BLASTP1.4 格式的记分矩阵文件。
-w #	相似性分值和序列比对输出行宽
-W #	比对周围序列正文的数量。缺省值为 30 个残基(fastx/y3，tfastx/y3)
-x "#，#"	比对计数的偏移查询和文库序列
-z #	指定统计学计算。缺省值为-z 1(表 10.13)
-Z #	指定用于统计学显著性估计的文库大小

表 10.11　程序特异的命令行选项

fasta3，fastx/y3，tfastx/y3，tfasta3 选项	
-1	按“init1”分值归类
-3	(仅 tfasta3、tfastx3、tfasty3)使用前框翻译结果
-A	采用 Smith-Waterman 比对作为输出，Smith-Waterman 是蛋白质序列、fastx/y3 和 tfastx/y3 的缺省方式，但不是 tfasta3 或 fasta3 进行的 DNA 比较
-c #	区带优化(band optimization)阈值
-f #	间隙中第一个残基的罚分
-g #	间隙中其余残基的罚分
-h #	仅 fastx/y3、tfastx/y3——密码子间的框移罚分
-j #	仅 fasty3、tfasty3——密码子的替换罚分
-t #	翻译表——fastx/y3、tfastx/y3 和 tfasta3 现在支持 BLAST 翻译表。见 `http://www.ncbi.nlm.nih.gov/htbin-post/Taxonomy/wprintgc?mode=c/`
-y #	区带优化的宽度；DNA 和蛋白质的缺省值为 16，且 *ktup* = 2；蛋白质为 32 且 *ktup* = 1
ssearch3 命令行选项	
-f #	间隙中第一个残基的罚分
-g #	间隙中其余残基的罚分

当使用从弗吉尼亚大学分发的 FASTA 程序时，命令行选择必须放在其他的程序参数之前。FASTA 程序的标准调用是：

```
program -opt1 -opt2 arg2 query_file library ktup-opt
```

具体地说：

```
fasta3 -q -f -14 -w 75 -L -m 1 mgstm1.aa/slib/swiss 1
```

在后者，fasta3 程序运行于安静模式(`-q`)，间隙中第一残基的罚分是–14(`-f -14` 而非缺省的–12)，比对是以 75 个残基/行(`-w 75`)列出，在比对中显示文库序列的长描述段(`-L`)，比对的字符高亮显示出差异而非相似性(`-m 1`)。图 10.4 显示了常规比对(图 10.4A)和命令行选项产生的比对(图 10.4B)之间的差异。

A

```
>>GTT1_MUSDO GLUTATHIONE S-TRANSFERASE 1 (EC 2.5.1.18) (C (208 aa)
 initn: 1229 init1: 1229 opt: 1230 Z-score: 1472.4 expect() 2.3e-75
Smith-Waterman score: 1230;  85.024% identity in 207 aa overlap

               10        20        30        40        50        60
gi|121 MVDFYYLPGSSPCRSVIMTAKAVGVELNKKLLNLQAGEHLKPEFLKINPQHTIPTLVDNG
        .::::::::.:::::.:::::.:.:::::::::::::::::::::::::::::::::.
GTT1_M  MDFYYLPGSAPCRSVLMTAKALGIELNKKLLNLQAGEHLKPEFLKINPQHTIPTLVDGD
                10        20        30        40        50
```

B

```
>>GTT1_MUSDO GLUTATHIONE S-TRANSFERASE 1 (EC 2.5.1.18) (CLASS-THETA).    (208 aa)
 initn: 1229 init1: 1229 opt: 1230 Z-score: 1615.1 expect() 2.6e-83
Smith-Waterman score: 1230;  85.024% identity in 207 aa overlap

               10        20        30        40        50        60        70
gi|121 MVDFYYLPGSSPCRSVIMTAKAVGVELNKKLLNLQAGEHLKPEFLKINPQHTIPTLVDNGFALWESRAIQVYLVE
        x        x     x     x x                                 xX         x
GTT1_M  MDFYYLPGSAPCRSVLMTAKALGIELNKKLLNLQAGEHLKPEFLKINPQHTIPTLVDGDFALWESRAIMVYLVE
                10        20        30        40        50        60        70
```

图 10.4　选择性输出格式

A. 使用默认程序参数的比对；B. 使用命令行选项的比对

命令行选项可以分为 5 个基本类别：计分参数选项、统计学选项、算法指定选项、文件说明选项和输出选项。

10.4.1 改变计分参数

FASTA3 软件包的所有的程序用 2 类计分参数计算序列比对：代矩阵和间隙罚分。缺省的计分矩阵、间隙罚分、*E* 值取值点和比较算法列于表 10.12。fasta3、ssearch3、fastx/y3 和 tfastx/y3 程序采用 BLOSUM50 计分矩阵[16]进行蛋白质序列(包括翻译过来的蛋白质序列)比较。其他蛋白质计分矩阵可以用`-s` 选项指定。包括 BLOSUM62(`-s BL62`)、BLOSUM80(`-s BL80`)，以及 PAM250(`-s P250`)和 PAM120(`-s P120`)[17,18]，还有低进化距离矩阵 MDM10(`-s M10`)和 MDM20(`-s M20`)[19]。此外，其他计分矩阵也可以通过在含有替换值文件(`-s matrix.file`)中提供文件名的方法使用。FASTA 程序 3 版本采用了和 blastp 程序相同的替代矩阵格式，和 BLAST 程序一起分发的 pam 程序能用于产生适当的格式化矩阵。

表 10.12 FASTA 程序的缺省值

程序	查询	文库	记分(-s)矩阵	间隙(-f，-g)罚分	框移(-h，-j)	−*E*()阈值点	比对
fasta3	蛋白质	蛋白质	BLOSUM50	−12/−2		10.0	Smith-Waterman
	DNA	DNA	+5/−4	−16/−4		2.0	band Smith-Waterman [a]
ssearch3	蛋白质	蛋白质	BLOSUM50	−12/−2		10.0	Smith-Waterman
	DNA	DNA	+5/−4	−16/−4		2.0	Smith-Waterman
fastx3	DNA(1 strand)	蛋白质	BLOSUM50	−15/−2	−20	5.0	Smith-Waterman [b]
fasty3	DNA(1 strand)	蛋白质	BLOSUM50	−15/−2	−20/−20	5.0	Smith-Waterman [b]
tfastx3	蛋白质	DNA	BLOSUM50	−15/−2	−20	5.0	Smith-Waterman [b]
tfasty3	蛋白质	DNA	BLOSUM50	−15/−2	−20/−20	5.0	Smith-Waterman [b]
fastf3	混合多肽	蛋白质	MDM20			5.0	
tfastf3	混合多肽	DNA	MDM10			5.0	

a. 参考文献[28]；b. 参考文献[15]

DNA 序列比较中，替代矩阵是给每一匹配记 5 分，错配−4 分(和可能的核苷酸匹配记 2 分，错配−1 分)。采用其他替代矩阵可以用`-s dna-matrix.file` 选项。

BLOSUM50 矩阵在识别极远相关序列时很有效(对长的紧密相关序列也很有效)。用短序列[18]或紧密相关序列时(如小鼠蛋白质相对于小鼠 ESTs)用“更浅(shallower)的”计分矩阵会更有效，如 MDM10 和 MDM20 适合于非常短序列中的

少量改变。

FASTA 程序的间隙罚分可以通过-f 和-g 选项来改变；-f 指定间隙的第一个残基罚分值，-g 指定其他残基的罚分值。还可以用表达式：$q+rk$，其中 q 是打开间隙的罚分，r 是间隙中的残基罚分(k 是间隙长度)。因此，`-f -12,-g -2`(缺省的蛋白质搜索)和：$q=10, r=2$ 等效。蛋白质替代矩阵，如 BLOSUM50 和 PAM250(以 1/3-bit 为单位[18])，在间隙罚分为−12/2 或−14/−2[20]时很有效，而计分矩阵如 BLOSUM62 和 PAM120(以 1/2-bit 为单位)，则在较低的起始残基罚分时(`-f -8`)有效。

正如更浅的替代矩阵可能更适合于比较密切相关序列(如哺乳动物)那样，高的间隙罚分可能也一样适合。采用间隙罚分为−20/−4 的 MDM20 计分矩阵能使程序以高期望值识别差异为 20%~40%左右的序列，但是程序也可能会错过其蛋白质序列低于 30%相同的同源物。

fastx3/tfastx3 和 fasty3/tfasty3 程序还提供了其他间隙参数。fastx3/tfastx3 采用-h 选项指定框移(因为 fastx3 算法的性质决定这是一定会发生在两个密码子之间)的罚分值。fasty3/tfasty3 采用-h 选项设置密码子间的框移罚分，-j 指定密码子内的框移罚分值。当搜索含有 5%错误的 EST 序列时，用缺省值-h −20 和-j −20 就很好[15]。但是如果 DNA 序列是没有什么错误，则应该用更高的框移罚分值，因为这样能减小框外比对的噪声。

一般地说，FASTA 程序提供的缺省间隙参数是位于有用范围的低端。减小间隙罚分会引起比对冲局部向全局漂移。第一个残基(-f)的罚分微小增量有时会稍微改善比对的期望值，但研究者应该对随间隙罚分剧烈改变的边界计分十分谨慎。替代矩阵中的变化通常比间隙罚分中的小变化对结果的影响更大；用 PAM250 矩阵搜索的期望值通常是 10^{-3}~10^{-10} 比 BLOSUM50 低。例如，在表 10.7 中显示的分值，采用 PAM250 矩阵时，`gtt1_drome` 和 `xuzm32`，`xuzm31` 以及 `xuzm1` 比对的期望值分别从 8.5×10^{-8}、2.5×10^{-6}和 8.8×10^{-5}，下降至 7.1×10^{-5}、0.001 和 0.15。用 Monte-Carlo prss3 程序评价比对的显著性时，应该确认采用了相同的替代矩阵和间隙罚分。

10.4.2 统计学估计的多种选择

FASTA3 软件包的有用功能之一是能精确估计局部相似性分数的统计学意义，从蛋白质：蛋白质、DNA：DNA，或蛋白质：DNA 译文之间的比对都可以进行。FASTA3 软件包中的程序是基于非相关序列的分数分布的基础上来计算期望值。因此，对于典型的针对含有数万条非相关序列的搜索其统计学估计很精确，但如果数据库中不含非相关序列时就不太准了。FASTA3 程序提供 6 种统计学估计选项(表 10.13，参考文献[6])。-z 3 选项很有用，因为它在搜索不含非相关序列的数据库时，或仅比较一对序列时就可以使用。

表 10.13　统计学选项

选项	说明
-z1	无统计学估计，有时数据库中没有不相关序列时需要
-z 0	非比例的统计学估计，估计值是从均数和序列相似性分值计算出的，典型的使用情况是所有的文库序列长度差不多时
-z 1	回归-比例估计，相似性分值的均数和误差从取对数[log(n)]后计算得出
-z 2	log-校正估计，仅用于历史目的；此方法已经过时，不应该再用
-z 3	Altschul-Gish 估计(仅蛋白质)，不是从数据中估计参数，而是用 Altschul 和 Gish[29]发表的预先计算好的参数。-z 3 是当所搜索库中不相关序列不占主要部分时估计比对的显著性的唯一选项
-z 4	-z 1 的另一个选择，在参数估计时以一种不同的算法去除高分值的可能相关序列
-z 5	-z 1 的另一个选择，也采用分值误差(库序列长度)的 log(n)回归。而-z5 所提供的估计会更准确一些，也对数据的问题更敏感，尤其是对小库(<500 条数据)搜索时

统计学意义对数据库大小的依赖能使对不同数据库进行搜索的结果复杂化。-Z number 选项能用于迫使程序认为所搜索的数据库大小是“number”值，例如-Z 100 000 可使程序认为数据库中有约 100 000 个哺乳动物基因。(“number”取值不能小于所搜索数据库的真正大小。)这个选项和-z 3 选项并用在搜索一个小预选序列库时特别有用。

10.4.3 输入选择

FASTA 程序提供了许多改变查询序列的使用方式和选择数据库选项(表格 10.14)。最常使用的输入选项是-i，它使 DNA 搜索采用查询序列的反向互补形式。(和 FASTA 的 BLASTN 和 GCG 版本不同，弗吉尼亚大学 FASTA 程序在 1998 年 12 月的 FASTA 32 版本前没有此选项，当查询 DNA 序列时自动用前向和反向 DNA 链搜索。)

表 10.14　输入选项

选项	说明
@	除文件名外,FASTA3 程序还能接受 UNIX 和 Windows 计算机中的 stdin 文件流中的查询序列。此时，所有信息必须按命令行输入，如 fasta3-q @ /slib/swiss.seq 1 < query. aa 表示输入来自 stdin(<query. aa)，swiss. seq 文库将以 *ktup* = 1 搜索。@选项在 Web 服务器上的 perl 脚本中最常见
:#-#	指定一个子序列，查询序列文件名后可以跟一个“:”以及一个数值范围以指定一段序列。如果第一个数值未给出，就认定为 1。如果最后一个数值未给出，子序列就扩展到序列的末端。因此，gtt1_drome. aa:51~150 就是指定了自 51 个开始的 100 个残基。子序列范围可以在命令行或程序提示输入查询序列时给定。也可以在一个“@”(stdin)符号后给出。序列范围只能用于第一个(查询)序列
-i	(仅 DNA 查询)用查询序列的反向互补形式搜索
-l file	认定用于定位序列数据库的 FASTLIBS 文件
-n	规定输入(查询)序列作为 DNA 输入(仅 fasta3 和 ssearch3)
-N #	以“#”长度块读入长文库序列(如细菌基因组)；如-N 5000 按 5000 个残基为一组读入长序列
-q/Q	沉默。即不提示输入

FASTA 程序可以容易地定义仅使用查询序列的一部分进行搜索，方法是用“:”修饰查询序列的文件名。命令：

```
fasta3  gtt1_drome.aa:1-100 s
```

是让程序用查询序列 gtt1_drome 的前 100 个残基搜索由“s”缩写指定的数据库。

fasta3 和 ssearch3 采用一个简单的算法决定查询序列是蛋白质还是 DNA：如果序列 A+C+G+T 多于 85%，它被认定是 DNA，否则它被当作蛋白质序列。-n 选项强迫查询序列当作 DNA；-n 选项需要 DNA 序列通过 stdin(@)选项(表 10.14)来提供。不同于 BLAST 程序，FASTA 程序目前仅在查询序列和文库序列之间报告最好的比对，甚至当文库序列很长并且可能包含几百个基因时也是这样。缺省的 FASTA 把长 DNA 序列拆散为长度约 80 000 核苷酸的块，但是这对于基因密集的细菌、酵母和果蝇染色体来说，这个尺寸太大了。-N 5000 选项告诉 fasta3 和 tfastx/y3 程序以 5000 核苷酸的块读入长 DNA 序列。当扫描大的且基因稠密的 DNA 序列时很必要。

10.4.4 改变输出的外观

许多 FASTA 命令行选项可以改变比对输出外观(表 10.15 和表 10.16)。有改变

表 10.15 输出选项

-a	(仅适用于 fasta3 和 ssearch3)完整显示查询和文库序列，而非仅仅比对部分
-A	(仅适用于 fasta3 DNA)fasta3 可以对蛋白质序列(以及翻译的 fastx/y3 和 tfastx/y3)进行完全 Smith-Waterman[22]比对，对于 DNA 仅仅进行 DNA：DNA 的区带限制的比对。-A 选项使 fasta3 对 DNA 序列进行完全 Smith-Waterman 比对，但当序列较长时运算会很慢
-b #	要显示的高分值文库序列数目
-d #	要显示的高分值比对数目
-E #	显示分数和比对期望值[*E*]的阈值点(cutoff)。缺省时，蛋白质：蛋白质比较为-*E* 10；翻译的 DNA：蛋白质比较为-*E* 5；DNA：DNA 比较为-*E* 2。-*E* 阈值点级别高于-b 和-d 选项。例如，要显示至少 20 个分值和 5 个比对，选项应为：-*E* 1000.0 -b 20 -d 10
-F #	一个 *E* 值低限阈值点以防止非常相似的序列被显示。-F le-4 能防止程序显示库中 $E<10^{-4}$ 的序列。这个选项对有很多相近同源序列的大蛋白质家族进行远缘同源性分析时很有用
-H	不显示直方图
-L	对比对提供长序列描述。有些序列文库格式(尤其是重新格式化的 GCG 库)在正式的序列描述前包含了许多非信息文本，用此选项，则所有的序列描述都被显示
-m #	见表 10.16
-O file	把结果写入到文件 file 中。UNIX 和 Windows 用户应该用“>file”方法进行输出
-R file	把所有序列的直接结果写入 file 文件
-w #	比对输出的宽度。FASTA 程序的缺省显示是每行 60 个残基；此宽度可以用-w 选项改成 200 个残基
-W #	序列正文的总数。fasta3 和 ssearch3 提供比对中的邻近序列正文(fastx/y3 和 tfastx/y3 没有)。典型的正文数是输出行的一半，可以用-W 选项改变
-x “##”	序列坐标。正常情况下，FASTA 程序假设每一序列起始于残基 1。有时，采用不同的起始坐标很有用，如当比较 cDNA 和编码基因或仅仅是序列的一部分时。-x “1，-751”会告诉 fasta3 从库中序列的“-751”位开始，而非“1”。在 UNIX、DOS 和 Macintosh 系统中，两个数字必须由双引号(“…”)括上

显示在比对行上残基数目的选项、改变残基数目的选项和改变比对格式的选项。有 2 个选项很特别(表 10.16)：-m5 提供了序列比对以及比对部分和查询序列相比的粗略示意图。此图形使我们更易于快速察看查询序列以及不同的文库序列比对部分，这样可以突出显示查询序列的不同区域。-m6 选项和-m5 完全相同，但提供了 HTML 标出命令和链接到 Entrez 和其他网站进行研究，以确认其与文库序列的关系。

表 10.16 比对选项

-m 0	用“:”指明完全相同的比对残基，保守替换用“.”
-m 1	不标明相同部分，用“x”标明保守替换部分
-m 2	用“.”标明相同部分，用残基表示不相同部分
-m 3	比对按两条 fasta 格式序列打印出，间隙用“-”表示。这些文件有时作为其他程序输入时很有用
-m 4	不显示比对；而在查询序列上显示比对区域的比例图形(------)。在标明蛋白质的不同区域很有用
-m 5	-m 4 和-m 0 组合同时显示比例图形和比对
-m 6	和-m 5 相似，但是包含网页浏览器，如 Netscape 或 Internet Explorer 的 HTML 命令，并且还有简化查询数据库序列和搜索数据库的链接
-m 10	分栏输出格式，可被其他计算机程序读出。每一比对都有一系列标记标签，标明起始、结尾、分数、搜索参数和其他信息

10.5 在序列同源区之外——鉴定新的平行进化同源基因

使用 FASTA 和 BLAST 程序鉴定远缘序列已经有充分的综述[3~5]，所以在本章的最后一部分我们要考虑一个稍微不同的问题，即充分利用 FASTA 程序的灵活性与其高质量的比对。

这里，我们先从 EST 数据库中鉴定出已知的人或小鼠家族的新平行进化同源基因。例如，已知在人、小鼠、大鼠和其他哺乳动物中有两种人类前列腺素合成酶，COX1(pgh1_human)和 COX2(pgh2_human)。前列腺素合成酶是非甾体类抗炎药的作用靶点，包括阿司匹林和布洛芬。因此，找到此家族的其余成员很有意义，并且很可能其他前列腺素合成酶已经测序，有可能是大规模的 EST 测序或是基因组测序结果。

10.5.1 总体策略

平行进化同源基因是一个基因家族的多个成员(相关或同源)，在基因复制过程中和家族中的其他成员不同。定向进化同源基因的不同之处在于它们来自不同

种类，对 SwissProt 数据库搜索(表 10.17)显示有 2 个前列腺素合成酶(PGH)亚家族，但也找出了远缘的过氧化酶。人类 PGH1 和 PGH2 同工酶共有大约 65 的相同序列 [$E<10^{-165}$]。相反，人和小鼠的定向进化同源的 PGH1 序列共有 89.3%相同序列。我们期望新的人 PGH 合成酶和 PGH1/PGH2 之间有很强的相似性[$E<10^{-20}$]，但是却低于 80%。因为我们要扫描 EST 数据库以发现新的平行进化同源基因，我们期望那些相似性大于 90%~95%的序列来自编码已知蛋白质的 mRNAs，这些蛋白质有测序错误，而那些有 50%~90%相似性的序列是可能的平行进化同源基因。

表 10.17　前列腺素合成酶搜索结果

最好的分值:		长度	*E*(74 357)
PGH1_HUMAN	前列腺素 G/H 合成酶 1	599	3.9e−264
PGH1_SHEEP	前列腺素 G/H 合成酶 1	600	2.3e−244
PGH1_MOUSE	前列腺素 G/H 合成酶 1	602	9.5e−237
PGH2_CHICK	前列腺素 G/H 合成酶 2	603	1.2e−168
PGH2_HUMAN	前列腺素 G/H 合成酶 2	604	1.9e−165
PGH2_MOUSE	前列腺素 G/H 合成酶 2	604	2.4e−164
PGH2_CAVPO	前列腺素 G/H 合成酶 2	604	1.7e−163
PGH2_RAT	前列腺素 G/H 合成酶 2	604	1.4e−162
PERM_MOUSE	髓过氧物酶前体	718	0.0001
PERO_DROME	过氧物酶前体	690	0.00024
PERT_HUMAN	甲状腺过氧物酶前体	933	0.0003
PERM_HUMAN	髓过氧物酶前体	745	0.00034
PERT_PIG	甲状腺过氧物酶前体	926	0.0029
PERL_BOVIN	乳过氧物酶前体	712	0.016
PERT_MOUSE	甲状腺过氧物酶前体	914	0.02
PERL_HUMAN	乳过氧物酶 LPO	324	0.027
PERT_RAT	甲状腺过氧物酶前体	914	0.089
FBP1_STRPU	fibropellin I prec.	1064	0.16
PGCN_RAT	neurocan core prot. prec.	1257	0.21
FBP3_STRPU	fibropellin C prec.	570	0.31
PGCN_MOUSE	neurocan core prot. prec.	1268	0.33
PERE_MOUSE	eosinophil peroxidase prec.	716	0.51
NOTC_DROME	neurogenic locus notch prot.	2703	0.74
DLK_MOUSE	δ样蛋白质前体	385	0.86
PERE_HUMAN	嗜曙红细胞过氧化物前体	715	0.92
NTC1_MOUSE	neurogenic locus notch homolog	2531	0.94

用 fasta3(*ktup* = 2)在 SwissProt 数据库中搜索 pgh1_human 的结果

为鉴别出新的 pgh1_human 同形物，我们要用 fasty3 程序在人 EST 数据库(得

自 ftp://ncbi.nlm.nih.gov/blast/db/)中搜索蛋白质序列 pgh1_human 和 pgh2_human。选用 tfasty3 的理由：①我们想把一条蛋白质序列和 DNA(EST) 数据库比较；②将采用期望值 *E* 和百分比特征以鉴别匹配，所以需要高质量的蛋白质:DNA 比对(tfastx3 更快捷但是产生的比对质量较差，见参考文献[15])。然后我们将检查 EST 序列中与查询序列有显著相似性的序列，并把它们归类到 pgh1_human 和 pgh2_human 的定向进化同源基因或新的平行进化同源基因。

10.5.2 统计学意义和百分比特征

然而我们的目标是鉴定出相似的序列，而非与已知的前列腺合成酶一样的序列，通常采用的相似性标准(*E* 值和百分比特征)并不完全包括我们需要的信息。正如 pgh1_human 和 pgh2_human tfasty3 的搜索结果证明的那样(表 10.18)。和查询序列有较高序列相似程度的 EST 序列不一定具有更好的 *E* 值。

表 10.18 前列腺素合成酶 ESTs

pgh1_human:		长度	[f/r]	opt	$E(10^6)$	%ident.	I/II
gb|R96180	Pineal_gland_N3HPG	355	[f]	654	3e−38	98.0	I
gb|AA296431	脐静脉内皮	279	[f]	380	6.7e−19	59.1	II
gb|T29235	人骨骼	257	[f]	358	2.2e−17	63.3	II
gb|AA037294	衰老成纤维细胞_NbHSF	471	[f]	304	3.1e−13	98.0	I
gb|AI022012	衰老成纤维细胞_NbHSF	537	[r]	248	3.5e−09	64.5	II
gb|N79146	多发性硬化症_2NbHMSP	544	[f]	207	2.9e−06	100.0	I
gb|AA223896	NT2 神经元前体	97	[f]	185	1.3e−05	80.0	??
gb|AA485017	NCI_CGAP_GCB1	208	[f]	124	0.72	66.1	
pgh2_human:							
gb|AA296431	脐静脉内皮	279	[f]	574	1.4e−35	96.8	II
gb|T29235	人骨骼	257	[f]	536	1e−32	92.9	II
gb|AI022012	衰老成纤维细胞_NbHSF	537	[r]	541	1.1e−32	95.8	II
gb|R96180	Pineal_gland_N3HPG	355	[f]	410	6.3e−23	65.8	I
gb|AA223896	NT2 神经元前体	97	[f]	136	0.01	50.0	??
gb|AA885610	NCI_CGAP_Lu5	320	[f]	141	0.018	46.3	
gb|AA911293	NCI_CGAP_Lu5	172	[f]	131	0.049	43.6	

用 tfasty3 程序在 BLAST est-人类数据库中搜索 pgh1_human 和 pgh2_human 的结果，采用缺省的 BLOSUM50 计分模型。pgh1(COXI)或 pgh2(COXII)定向平行在右列中注明

E 值和百分比特征之间的分歧反映了 *E* 值依赖于比对序列的长度。EST 序列比较不完整，所以一条和 gb|N79146 C 端有 30 个氨基酸完全相同的定向进化同源基因的期望值(2.9×10^{-6})比和一条 paralogous 基因有 59%相同时的期望值

[$E<6.7\times10^{-19}$]更差。而相同的百分率是一个更差的相似性标准。因为无关序列(如 gb|AA485017)可能有更多部分相同(62 个密码子中有 66.1%相同)，但这不能形成有统计学意义的相似性分数。无论如何，有显著相似性的序列之间，相同百分率是序列差异的有用度量。因此，在表 10.18 中有统计学意义的匹配中，orthologous 匹配的相同百分率总是>90%，只有一个可能是例外(gb|AA223896，见 10.5.3 节)。

10.5.3 用计分模型移动进化水平线

表 10.18 中用 pgh1_human 和 pgh2_human 找到的高分值 ESTs 提示所有这样的 ESTs(一个例外)都和 pgh1_human 与 pgh2_human 有 90%以上的残基相同。那个例外——gb|AA223896 也和 pgh1_human 有 80%相同，和 pgh2_human 有 50%相同，因此是一候选的新前列腺素合成酶平行进化同源基因。

但是，gb|AA223896 EST 序列太短(97 核苷酸)，仅有 6 个错配，半数都在序列末端的 20 个核苷酸内。因此，这是否是一个新平行进化同源基因，或者是一条短的、低质量的 pgh1_human mRNA 序列，末端有几个错误(和所期望的高 EST 序列通常一样)。但是，末端错误问题可以通过对比对代码的特定改变使末端-错配的影响变小，更简单的办法就是采用更浅的计分模型。

用很浅的计分模型(MDM20 和 MDM10，参考文献[19]，表 10.19)进行搜索后

表 10.19 浅计分模型搜索

pgh1_human:	长度	E(BL50)	%	E(M20)	%	E(M10)	%	I/II
gb|R96180	355	3e–38	98.0	2.3e–72	99.0	6.5e–75	100.0	I
gb|AA296431	279	6.7e–19	59.1	6.8e–25	61.3	1.3e–22	62.4	II
gb|T29235	257	2.2e–17	63.3	5.3e–22	64.8	2.6e–18	66.2	II
gb|AA037294	471	3.1e–13	98.0	3e–30	98.0	3.3e–31	97.8	I
gb|AI022012	537	3.5e–09	64.5	1.2e–15	58.8	3.4e–13	60.8	II
gb|N79146	544	2.9e–06	100.0	2.6e–16	100.0	3.0e–17	100.0	I
gb|AA223896	97	1.3e–05	80.0	8.4e–13	87.1	2.8e–12	87.1	??
gb|AA485017	208	0.72	66.1	4.8e–14	84.7	4.1e–14	88.9	??
pgh2_human:								
gb|AA296431	279	1.4e–35	96.8	2.2e–69	96.8	8.0e–72	98.9	II
gb|T29235	257	1e–32	92.9	2.9e–61	94.1	9.1e–63	95.2	II
gb|AI022012	537	1.1e–32	95.8	1.6e–68	96.0	1.1e–70	97.0	II
gb|R96180	355	6.3e–23	65.8	1.0e–30	56.9	9.1e–27	60.3	I
gb|AA485017	208	—[a]	–	2.4e–05	75.6	3.3e–4	79.1	??
gb|AA223896	97	0.01	50.0	0.01	69.0	0.2	79.2	??
gb|AA885610	320	0.018	46.3	—	—	—	—	
gb|AA911293	172	0.049	43.6	—	—	—	—	

a. E() 值表示为——时即>5.0

显示了略微的不同，蕴涵着更吸引人的前景。当采用更浅的计分模型时，定向进化同源基因和平行进化同源基因二者的比对变得更具有统计学意义，而且，像所期望的那样，序列相同的百分率增加了(更浅的模型对相同部分打分更多，而对非保守替换扣分更多)。最有趣的是两条序列 gb|AA223896 和 gb|AA485017，对 pgh1_human 采用 MDM20 和 MDM10 时显示了显著的相似性。两条序列都是新平行进化同源基因(因为采用 MDM20 时定向进化同源基因总是有高于 90%的相同率)，但是两条序列比对后显示很多框移(不影响相同百分率计算)，提示可能因为序列质量差使这些序列相同率<90%，而非一个新的基因。

pgh2_human 中最后的两条记录(gb|AA885610 和 gb|AA911293)的搜索显示浅计分模型也能被用于快速选出高分无关序列。这两条序列的期望值在采用 BLOSUM50 模型时很有意义(0.018 和 0.049)，和 pgh1_human 无显著相似性，但采用 MDM20 和 MDM10 时变得很高[E>5]。因此，当期望查询序列中的至少一条近乎相同时，更浅的计分模型可以用于提供一个更严谨的序列相似性测验。而且 MDM20 和 MDM10 可以提供更严谨的比对，它们不是最好的模型，因为它们是为假设的进化模型而建的。更精确的模型可以从观察大量的 EST 序列错误后进行修改得到，这样把模型建立在测序错误模型上，而非进化趋异。

10.6 总结

FASTA3 和 FASTA2 软件包提供了一套灵活的序列比较程序，这套程序非常有价值，因为它们提供精确的统计估计和高质量的比对。传统上，序列相似性搜索寻求解答一个问题：“待查询序列和数据库中的序列有同源性吗？”FASTA 和 BLAST 程序都能以其统计学估计提供可靠的答案：如果期望值 E<0.001~0.01，并且你一天不做上百个搜索，那么答案就可能是“是”。

一般来说，最有效的搜索策略遵循下列规则：

(1) 如果可能，比较氨基酸水平，而不是核苷酸水平。先用蛋白质序列进行搜索(blastp，fasta3 和 ssearch3)，然后用翻译的 DNA 序列(fastx，blastx)，DNA 水平只是最后的手段(表 10.5)。

(2) 先搜索可能含有感兴趣序列的最小数据库(但是必须含有许多无关序列以进行准确的统计学估计)。

(3) 采用序列统计学，而非相同百分率和相似百分率，作为你序列同源性的首要标准。

(4) 通过查找无关序列的最高得分来核对统计学是否精确。采用 prss3 确认期望值，用查询序列的乱序拷贝进行搜索[randseq，以打乱的序列进行搜索其 E 值应该约为 1.0]。

(5) 考虑以不同的间距罚分和其他计分模型进行搜索。当采用 BLOSUM62 而非 BLOSUM50，或者用–14/–2 的间距罚分代替–12/–2 时，进行长查询序列对全长序列库搜索将不会剧烈变化[20]。但是，更浅或更严谨的计分模型在部分序列的剥离关系时更有效[3, 18]，并且他们能使相似性搜索的范围迅速缩小。

但是，正如最后部分所阐明的那样，*E* 值仅为描述一条序列关系的第一步。一旦确认序列是同源的，随后应该看一下序列比对和相同百分率，尤其是用低质量序列进行搜索时。当序列比对非常短时，采用更浅的模型时比对应该变得更显著，例如，BLOSUM62 与 BOLOSU50 对比(记得改变间距罚分)。

同源性能从统计学显著的相似性中可靠地推导出来。而同源性意味着共同的三维结构，不一定意味着相同的功能。定向进化同源基因序列通常有相同的功能，但是平行进化同源基因序列常常功能不同。基序数据库，如 PROSITE[21]，能提供关键功能残基保守的证据。然而，在缺乏全部序列上的相似性时，基序识别不是同源性的可靠指标。

致谢

W. R. P. 由国家医学图书馆提供资助(LM04961)。

(阮承迈　译)

参　考　文　献

[1] Wilbur, W. J. and Lipman, D. J. (1983) Rapid similarity searches of nucleic acid and protein data banks. *Proc. Natl. Acad. Sci. USA* **80,** 726-730.

[2] Bult, C. J., White, O., Olsen, G. J., Zhou, L., Fleischmann, R. D., Sulton, G. G., Blake J. A., Fitzgerald, L. M., Clayton, R. A., Gocayne, J. D., Kerlavage, A. R., Dougherty, B. A., Tomb, J.-F., Adams, M. D., Reisch, C. I., Overbeek, R., Kirkness, E. F., Weinstock, K. G., Merrick, J. M., Glodek, A., Scott, J. L., Geoghagen, N. S. M., Weidman, J. F., Fuhrmann, J. L., Nguyen, D., Utterback, T. R., Kelley, J. M., Peterson, J. D., Sadow, P. W., Hanna, M. C., Cotton, M. D., Roberts, K. M., Hurst, M. A., Kaine, B. P., Borodovsky, M., Klenk, H.-P., Fraser, C. M., Smith, H. O., Woese, C. R., and Venter, J. C. (1996) Complete genome sequence of the methanogenic archaeon, methanococcus jannaschii. *Science* **273,** 1058-1073.

[3] Altschul, S. F., Boguski, M. S., Gish, W., and Wootton, J. C. (1994) Issues in searching molecular sequence databases. *Nat. Genet.* **6,** 119-129.

[4] Pearson, W. R. (1996) Effective protein sequence comparison. *Meth. Enzymol.* **266,** 227-258.

[5] Pearson, W. R. (1997) Identifying distantly related protein sequences. *Comput. Appl. Biosci.* (now *Bioinformatics*) **13,** 325-332.

[6] Pearson, W. R. (1998) Empirical statistical estimates for sequence similarity searches. *J. Mol. Biol.* **276,** 71-84.

[7] Pearson, W. R. and Lipman, D. J. (1988) Improved tools for biological sequence comparison *Proc. Natl. Acad. Sci. USA* **85,** 2444-2448.

[8] Lipman, D. J. and Pearson, W. R. (1985) Rapid and sensitive protein similarity searches. *Science* **227,** 1435-1441.

[9] Bleasby, A. J., Akrigg, D., and Attwood, T. K. (1994) Owl-a non-redundant composite protein sequence database. *Nucleic Acids Res.* **22,** 3574-3577.

[10] Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990) A basic local alignment search tool. *J. Mol. Biol.* **215,** 403-410.

[11] Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997) Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic Acids Res.* **25,** 3389-3402.

[12] Arratia, R., Gordon, L., and Waterman, M. S. (1986) An extreme value theory for sequence matching. *Ann. Stat.* **14,** 971-993.

[13] Karlin, S. and Altschul, S. F. (1990) Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc. Natl. Acad. Sci. USA* **87,** 2264-2268.

[14] Wootton, J. C. and Federhen, S. (1993) Statistics of local complexity in amino acid sequences and sequence databases. *Comput. Chem.* **17,** 149-163.

[15] Pearson, W. R., Wood, T., Zhang, Z., and Miller, W. (1997) Comparison of DNA sequences with protein sequences. *Genomics* **46,** 24-36.

[16] Henikoff, S. and Henikoff, J. G. (1992) Amino acid substitutions matrices from protein blocks. *Proc. Natl. Acad. Sci. USA* **89,** 10 915-10 919.

[17] Schwartz, R. M. and Dayhoff, M. (1978) Matrices for detecting distant relationships, in *Atlas of Protein Sequence and Structure,* vol. 5, suppl. 3 (Dayhoff, M., ed.) National Biomedical Research Foundation, Silver Spring, MD, pp. 353-358.

[18] Altschul, S. F. (1991) Amino acid substitution matrices from an information theoretic perspective. *J. Mol. Biol.* **219,** 555-565.

[19] Jones, D. T., Taylor, W. R., and Thornton, J. M. (1992) The rapid generation of mutation data matrices from protein sequences. *Comp. Appl. Biosci.* (now *Bioinformatics*) **8,** 275-282.

[20] Pearson, W. R. (1995) Comparison of methods for searching protein sequence databases. *Protein Sci.* **4,** 1145-1160.

[21] Bairoch, A. (1991) PROSITE: a dictionary of sites and patterns in proteins. *Nucleic Acids Res.* **19** (suppl) 2241-2245.

[22] Smith, T. F. and Waterman, M. S. (1981) Identification of common molecular subsequences. *J. Mol. Biol.* **147,** 195-197.

[23] Huang, X. and Miller, W. (1991) A time-efficient, linear-space local similarity algorithm. *Adv. Appl. Math.* **12,** 337-357.

[24] Waterman, M. S. and Eggert, M. (1987) A new algorithm for best subsequences alignment with application to tRNA-rRNA comparisons. *J. Mol. Biol.* **197,** 723-728.

[25] Myers, E. W. and Miller, W. (1988) Optimal alignments in linear space. *Comp. Appl. Biosci.* **4,** 11-17.

[26] Kyte, J. and Doolittle, R. F. (1982) A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.* **157,** 105-132.

[27] Barker, W. C., Garavelli, J. S., Haft, D. H., Hunt, L. T., Marzec, C. R., Orcutt, B. C., Srinivasarao, G. Y., Yeh, L. S. L., Ledley, R. S., Mewes, H. W., Pfeiffer, F., and Tsugita, A. (1998) The PIR-International protein sequence database. *Nucleic Acids Res.* **26,** 27-32.

[28] Chao, K.-M., Pearson, W. R., and Miller, W. (1992) Aligning two sequences within a specified diagonal band. *Comp. Appl. Biosci.* (now *Bioinformatics*) **8,** 481-487.

[29] Altschul, S. F. and Gish, W. (1996) Local alignment statistics. *Meth. Enzymol.* **266,** 460-480.

11 采用CLUSTAL W和CLUSTAL X进行多序列比对

Ashok Aiyar

11.1 引言

进行多个蛋白质和核酸序列的比对目的有两个：首先是找出序列中有保守生物学功能的共同的基序，其次是找出新发现序列中可能对了解其生物学功能有用的基序。通常是通过在数据库中扫描新发现序列来做到这一点。

CLUSTAL W 和 CLUSTAL X 是两个用于快速和可靠地比对多个蛋白质和核酸序列的相关程序。本章将介绍使用这些程序通过多序列比对找出蛋白质和核酸之间相同的序列模式和基序，以及阐明如何用这些程序把一个新序列和原先已经比对的一套序列或模式比对。两个程序都可以把比对好的一套序列构建进化树，因为还有其他程序和软件包能进行直接的进化树分析，这里就不对 CLUSTAL W 和 CLUSTAL X 的此项功能作详细介绍了。

CLUSTAL 原先是为运行 MS DOS 的 IBM PC-兼容型计算机所编写的[1]。后来，一个重新编写的版本 CLUSTAL V，成为多种计算机和操作系统上的免费软件，并且带有源码[2]。两个最新的版本 CLUSTAL，CLUSTAL W 和 CLUSTAL X，对 CLUSTAL V 进行了一些改进，改善了多序列比对的可靠性和敏感性，而且运行速度不降低[3, 4]。CLUSTAL W 和 CLUSTAL X 的主要区别在于：前者有一个简单的文本界面，而后者采用国家生物技术信息中心(NCBI)的 VIBRANT 工具箱构建的优美的图形界面。CLUSTAL X 还提供给用户进行多序列和模式比对时的多种选项。二者最新的版本是 CLUSTAL W 版本 1.75 和 CLUSTALX 版本 1.65b(注：现在应该有更新的版本)。无论版本号有何不同，两个程序都有相同的匹配和多序列比对算法。CLUSTAL 程序已经为大多数操作系统预先编译好，可以直接运行在包括 DOS、Linux、MacOS，以及不同版本的 UNIX、VMS 和 Windows 95/NT 的操作系统上。全部源码都免费提供，并且可以很容易地用于其他操作系统上。ANSI C 兼容的编译器，如 GNU C(gcc)编译器重新编译。虽然本章中的所有例子都是在 Linux 中运行 CLUSTAL W 和 CLUSTAL X 得到的，但是，不用担心，程序提供的界面在其他操作系统中都很相似。本章中讨论两者都有的特点时，用 CLUSTAL(W/X)表示。如果是独有的特点，则用单一的程序名。

CLUSTAL(W/X)通过 Feng 和 Doolittle[5, 6]的多序列比对的改进算法来进行比对。比对分为 4 步，如图 11.1 所示：首先，对指定序列进行多序列配对比对。其次，这些配对比对用于计算相似性分数(相同百分比)，再采用 Neighbor Joining(NJ)算法做无根树[7]。这些树有分支，其距离每一分支节点的长度和估计的离散度成比例。第三步，这些无根树用中点算法转换成一个有根数[8]。在此步骤，有根树的分支长度被用来计算每一序列的权值，具体见 Thompson 等人的论文[3]。最后，为产生多序列比对，作为向导的有根树用于比对越来越大的序列组，从树梢向树根进行处理。在每一步骤中，同时使用一种动态编程算法和残基特异的权值矩阵，还有间隙-打开/间隙-扩展罚分(gap-opening/gap-extension)来进一步比对前边比对的序列[3,4]。

多种不同的残基-权值矩阵能用于蛋白质比对，包括一种简单的相似性矩阵，或者 BLOSUM[9]、PAM[10]和 GONNET[11]系列矩阵。至于核酸比对，用户可以在 IUB 矩阵或 CLUSTAL W[1,6]相似性矩阵[3]之间选择。

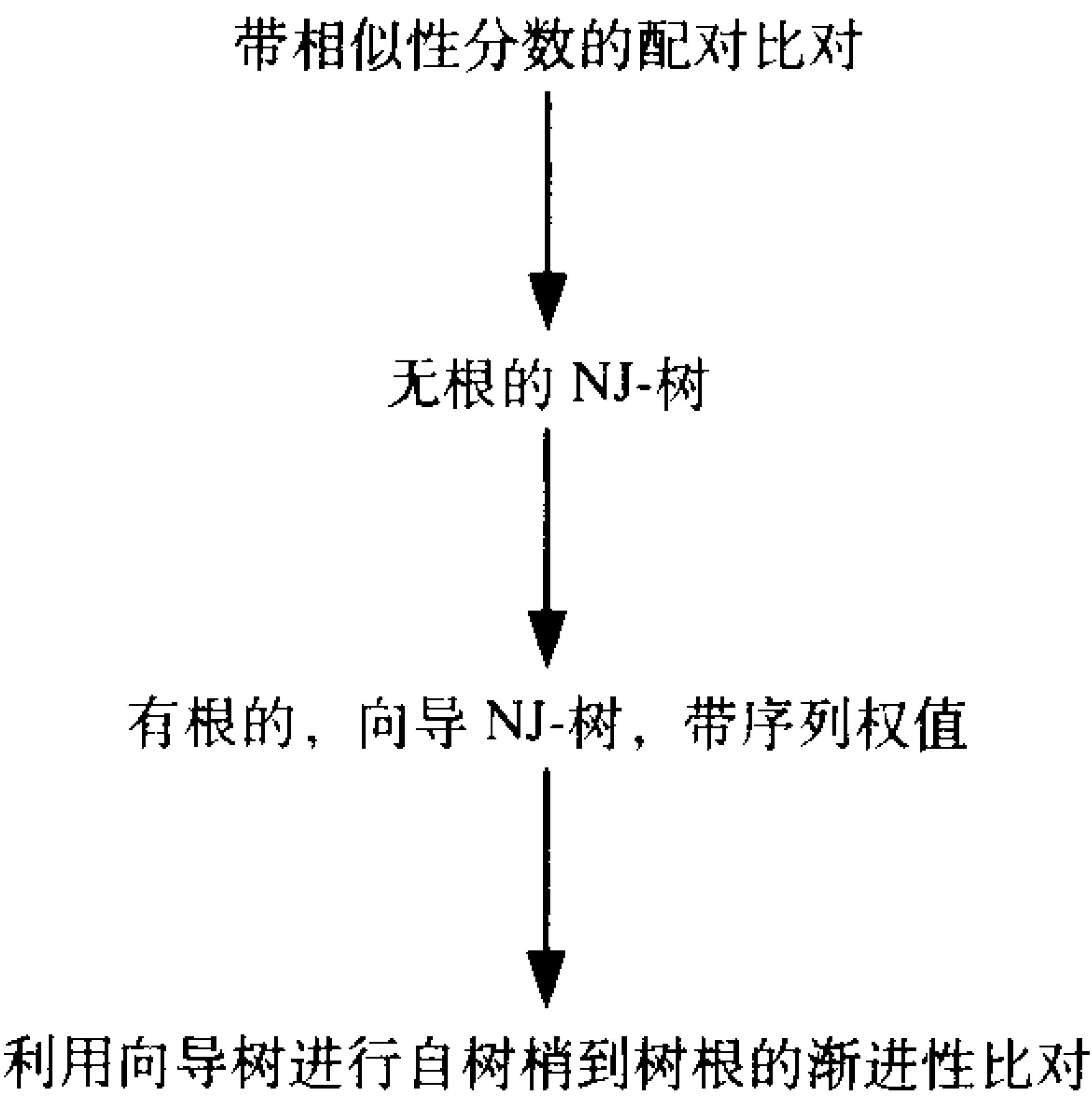


图 11.1　CLUSTAL(W/X)采用的渐进性比对算法概括

11.2　材料

11.2.1　获得 CLUSTAL(W/X)程序

CLUSTAL(W/X)程序和源码以及安装说明都免费提供。二者可以从下列 URL 得到：

CLUSTAL W：`ftp://ftp-igbmc.u-strasbg.fr/pub/ClustalW`
CLUSTAL X：`ftp://ftp-igbmc.u-strasbg.fr/pub/ClustalX`

此外，不同操作系统的 CLUSTAL(W/X)版本可以自下列网址得到：

UNIX/Linux:

ftp://ftp.ebi.ac.uk/pub/software/:>unix/clustalw

ftp://ftp.ebi.ac.uk/pub/software/:>unix/clustalw/clustalx

DOS,Windows 95/NT:

ftp://ftp.ebi.ac.uk/pub/software/dos/clustalw

ftp://ftp.ebi.ac.uk/pub/software/dos/clustalw/clustalx

http://www.csc.fi/molbio/progs/clustalw

MacOS:

ftp://ftp.ebi.ac.uk/pub/software/>mac/clustalw

ftp://ftp.ebi.ac.uk/pub/software/>mac/clustalw/clustalx

http://www.csc.fi/molbio/progs/clustalw

VMS:

ftp://ftp.ebi.ac.uk/pub/software/vax/clustalw

11.2.2 安装 CLUSTAL(W/X)

所提供的 CLUSTAL(W/X)程序应该足以应付大多数的需要。安装时，把程序拷贝到所有用户都能看到的目录下，如在 PATH 环境变量中的目录。在线帮助文件"clustalw_help"应该置于和可执行代码相同的目录下。对于 CLUSTAL X，则同样安装文件"clustalx_help"。此外，参数文件(*.par)也应该被拷贝到此目录中。

11.2.3 编译 CLUSTAL(W/X)

如果需要改变程序运行时不能设置的程序参数，或用于特殊目的的比对算法，那么你就要编译 CLUSTAL(W/X)的源码，方法见下述。

11.2.3.1 编译 CLUSTAL W

需准备一个 ANSI C 编译器，如 gcc 或 egcs 来编译 CLUSTAL W。

(1) 创建一个目录"clustalw"并拷贝分发的文档到其中。此文档应该是"clustal1.75.UNIX. tar. Z"。

(2) 执行解压缩命令：

```
cat clustalw1.75.UNIX.tar.Z|uncompress|tar xvf -
```

(3) 利用提供的 Makefile 进行编译，其与 GNU 兼容。

(4) 如 11.2.2 节所述安装编译好的程序。

11.2.3.2 编译 CLUSTAL X

和 CLUSTAL W 一样，需要 ANSI C 编译器来编译 CLUSTAL X。此外，还要编译和安装 NCBI VIBRANT 工具箱。

(1) Windows 95/NT、Macintosh 和 UNIX/Linux 版本的 NCBI 工具箱源码在：`ftp://ncbi.nlm.nih.gov/toolbox/ncbi_tools`。下载并安装适合你的操作系统的工具箱。在 UNIX/Linux 系统上，编译 NCBI 工具箱的 VIBRANT 库时需要 OSF/Motif 2.0 或更高版本。CLUSTAL X 需要在 VIBRANT 库的基础上编译，因此关键在于安装合适的 Motif 版本。如果 Motif 2.0 还未在计算机上安装，可以从操作系统(UNIX/Linux)的经销商那里，或者从单位的计算机服务部门得到。按 NCBI 工具箱包中所带的说明编译 NCBI 库。

(2) 解压缩 CLUSTAL X 到“clustalx”目录下，方法如 CLUSTAL W 中所述。编辑所提供的 Makefile 确定 NCBI 工具箱安装目录。

(3) 用编辑好的 Make 和 GNU make 编译 CLUSTAL X。

(4) 安装编译好的程序(如 11.2.2 节所述)。

11.3 方法

先在 11.3.1 节和 11.3.2 节中分别描述如何用 CLUSTAL W 和 CLUSTAL X 进行多项比对。然后在 11.3.3 节中描述用 CLUSTAL(W/X)进行工程文件比对。最后在 11.3.4 节中描述如何用 CLUSTAL(W/X)构建进化树。

当描述使用 CLUSTAL(W/X)创建多项比对时，将以牛免疫缺陷病毒(BIV)、牛白血病病毒(BLV)、马传染性贫血病毒(EIAV)、人艾滋病病毒(HIV1)、小鼠乳腺肿瘤病毒(MMTV)、Mason-Pfizer 猴病毒(MPMV)、羊 lentivirus(OLV)和罗氏肉瘤病毒(RSV)[12]的核壳体(nucleocapsid)蛋白(NC)为例。NC 是一种小 RNA-结合蛋白质，有 1~2 个锌结合基序 $CX_2CX_4HX_4C$，称为 cys-his 盒。除此以外，没有发现反病毒 NC 蛋白之间有其他保守序列[13, 14]。

11.3.1 用 CLUSTAL W 进行蛋白质序列的多序列比对

(1) 欲比对的序列应该在一个文件中。此文件必须是如下格式之一：EMBL/SwissProt[15], NBRF/PIR[16], FASTA[17], GCG/MSF[18], GCG/RSF[18], GDE[19]或 CLUSTAL。

(2) 执行程序，键入 `clustalw` 然后按 Enter(回车键)。如果程序不执行，则看看当前目录是否位于 clustalw 可执行程序目录下。运行后，可见到如图 11.2 所示菜单。其中在线帮助和许多联机帮助可以点击 H 选项，从该菜单和许多子菜单中获得。选项 1 可以加载包含序列的待排列文件。

(3) 一旦序列加载，可由选项 2 进入多项对比菜单。点击选项 2 可见图 11.3 所示的菜单。从该菜单点取选项 1 可以以程序缺省参数开始多序列比对，这些参数可通过这个菜单中的其他选项来修改，尤其是菜单中的 5、6 选项，可以分别定义用于一对和多对比对参数，如下面(5)、(6)项所述。

```
**************************************************************
******** CLUSTAL W (1.74) Multiple Sequence Alignments  ********
**************************************************************


    1. Sequence Input From Disc
    2. Multiple Alignments
    3. Profile / Structure Alignments
    4. Phylogenetic trees

    S. Execute a system command
    H. HELP
    X. EXIT (leave program)

Your choice: []
```

图 11.2　CLUSTAL W 主菜单

```
****** MULTIPLE ALIGNMENT MENU ******


    1.  Do complete multiple alignment now (Slow/Accurate)
    2.  Produce guide tree file only
    3.  Do alignment using old guide tree file

    4.  Toggle Slow/Fast pairwise alignments = SLOW

    5.  Pairwise alignment parameters
    6.  Multiple alignment parameters

    7.  Reset gaps before alignment? = OFF
    8.  Toggle screen display          = ON
    9.  Output format options

    S.  Execute a system command
    H.  HELP
    or press [RETURN] to go back to main menu

Your choice: []
```

图 11.3　CLUSTAL W 多序列比对菜单

(4) 选择选项 1 后，出现一个交互式菜单，允许用户确定输出的树文件和比对结果文件名。缺省状态下，系统会选择和输入文件同名，但扩展名分别为“.dnd”和“.aln”的文件名。多序列输出的文件名说明见下面的第(7)项。

(5) 匹配比对参数菜单(图 11.4)，允许用户设置慢速和快速的匹配比对参数。慢而精确的比对时，空位(间隙)罚分和邻接区域，以及残基的权值矩阵可以由用户自行定义。缺省时，蛋白序列用的是 BLOSUM 矩阵，而核酸用的是 IUB 矩阵。如果选择快速的匹配比对(图 11.3 中菜单的选项 4)，用户可以改变 k-tuple 大小、空位罚分、窗口大小和顶部对角线的数量。如果要求最大的敏感度，可采用小的 k-tuple 大小(如 word size)和更大的窗口；如果要求速度快则反之。

```
********* PAIRWISE ALIGNMENT PARAMETERS *********

       Slow/Accurate alignments:

       1. Gap Open Penalty        :10.00
       2. Gap Extension Penalty   :0.10
       3. Protein weight matrix   :BLOSUM series
       4. DNA weight matrix       :IUB

       Fast/Approximate alignments:

       5. Gap penalty             :3
       6. K-tuple (word) size     :1
       7. No. of top diagonals    :5
       8. Window size             :5

       9. Toggle Slow/Fast pairwise alignments = SLOW

       H. HELP

Enter number (or [RETURN] to exit): ▯
```

图 11.4　双序列对比参数菜单

(6) 在多序列比对参数菜单中，如图 11.5 所示，用户可以调节空位和邻接区域的罚分，以及所用的残基权值矩阵。在分析远缘相关序列时“delay divergent

```
********* MULTIPLE ALIGNMENT PARAMETERS *********

       1. Gap Opening Penalty              :10.00
       2. Gap Extension Penalty            :0.05
       3. Delay divergent sequences        :40 %

       4. DNA Transitions Weight           :0.50

       5. Protein weight matrix            :BLOSUM series
       6. DNA weight matrix                :IUB
       7. Use negative matrix              :OFF

       8. Protein Gap Parameters

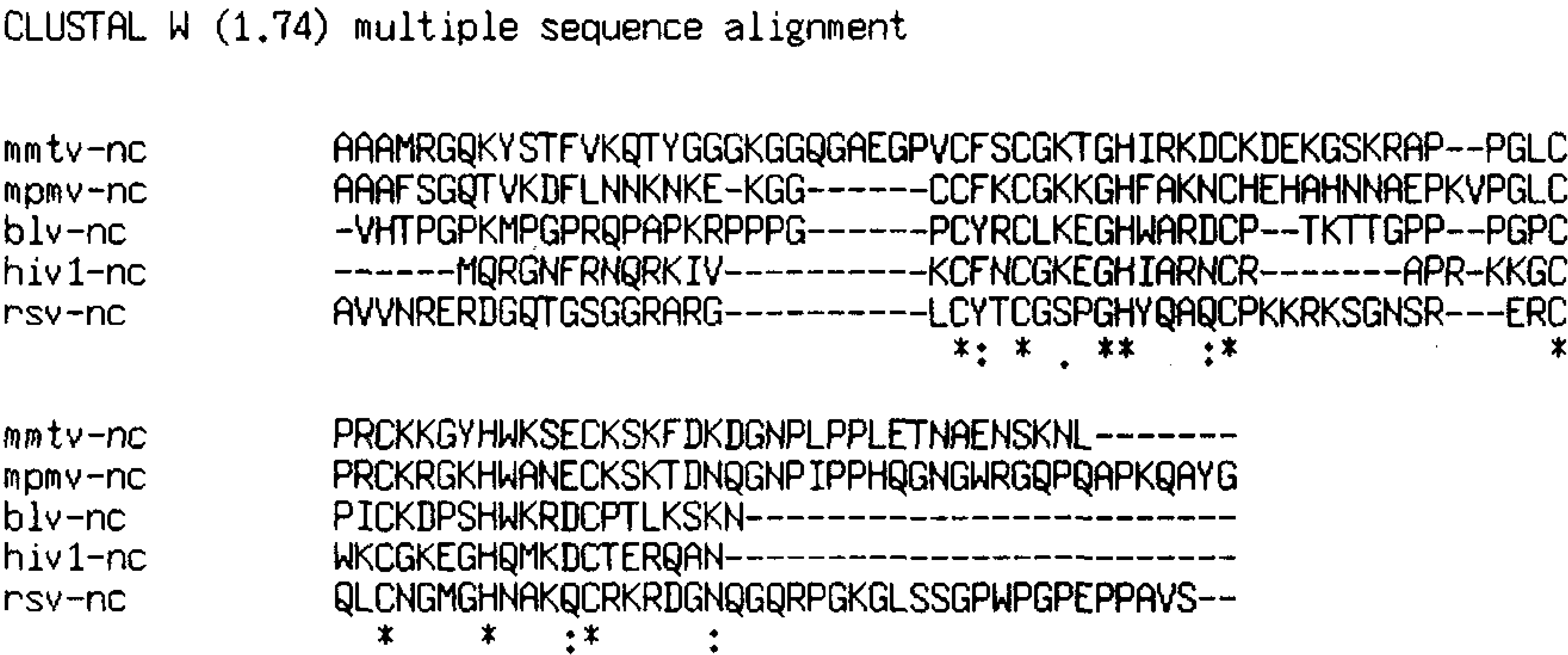
       H. HELP

Enter number (or [RETURN] to exit): █
```

图 11.5　多序列比对参数菜单

sequence”选项就特别有用。序列相似程度低于此类时就会稍后比对，以优化一致序列和空位的插入，这些是在比对早期采用所输入序列中发散程度低的序列之间进行的。

(7) 不同的输出格式可以用多项比对菜单的选项 9 来设置，如图 11.3 所示。缺省的比对格式是 CLUSTAL。图 11.6 是一个输出比对格式的例子，其中，BLV、HIV1、MMTV、MPMV 和 RSV 的核壳体蛋白用缺省的 CLUSTAL W 参数比对，图 11.6 的比对输出文件(.aln 文件)是一个 ASCII 文本文件，可以输入到比对编辑器或打印。所有五种 NC 蛋白的保守部分(cys-his 盒)以星号标示。比对输出格式也可以 GCG/MSF、NBRF/PIR、GDE 和 PHYLIP[20]格式保存。GCG/MSF 输出格式可用于 GCG 程序，如 PRETTY 和 PROFILEMAKE[18]输入。PHYLIP 输出格式可用于 PHYLIP 进化树分析软件包的比对结果输入[20,21]。



```
CLUSTAL W (1.74) multiple sequence alignment

mmtv-nc        AAAMRGGQKYSTFVKQTYGGGKGGQGAEGPVCFSCGKTGHIRKDCKDEKGSKRAP--PGLC
mpmv-nc        AAAFSGQTVKDFLNNKNKE-KGG------CCFKCGKKGHFAKNCHEHAHNNAEPKVPGLC
blv-nc         -VHTPGPKMPGPRQPAPKRPPPG------PCYRCLKEGHWARDCP--TKTTGPP--PGPC
hiv1-nc        ------MQRGNFRNQRKIV----------KCFNCGKEGHIARNCR-------APR-KKGC
rsv-nc         AVVNRERDGQTGSGGRARG----------LCYTCGSPGHYQAQCPKKRKSGNSR---ERC
                                             *: * . **  :*                  *

mmtv-nc        PRCKKGYHWKSECKSKFDKDGNPLPPLETNAENSKNL-------
mpmv-nc        PRCKRGKHWANECKSKTDNQGNPIPPHQGNGWRGQPQAPKQAYG
blv-nc         PICKDPSHWKRDCPTLKSKN------------------------
hiv1-nc        WKCGKEGHQMKDCTERQAN-------------------------
rsv-nc         QLCNGMGHNAKQCRKRDGNQGQRPGKGLSSGPWPGPEPPAVS--
                 *    *   :*     :
```

图 11.6　5 个反转录病毒核酸包壳蛋白多序列比对

11.3.2　采用 CLUSTAL X 进行多蛋白序列比对

(1) CLUSTAL X 具备 CLUSTAL W 的所有功能，并带有易用的图形用户界面，还有一些在 CLUSTAL W 中没有的特点，在本节中要涉及。在 UNIX 或 Linux 下执行 CLUSTAL X 需要 X-终端，而在 WINDOWS 95/NT 和 MacOS 版本中则不需要。

(2) CLUSTAL X 的主显示窗口是缺省加载的，但可以设置成预置比对模式(profile alignment mode)。在此模式中用户可以把新序列和原先比对好的序列之间比对，或比对两套比对结果。预置比对在 11.3.3 节中介绍。

(3) 启动 CLUSTAL X 后，在 File 菜单的 Load sequences 项加载要比对的序列文件。输入文件的格式限制和 11.3.1 节的(1)项相同。在加载序列文件后，显示如图 11.7 的窗口。和 CLUSTAL W 相似，Help 菜单有在线帮助。主窗口中序列顺

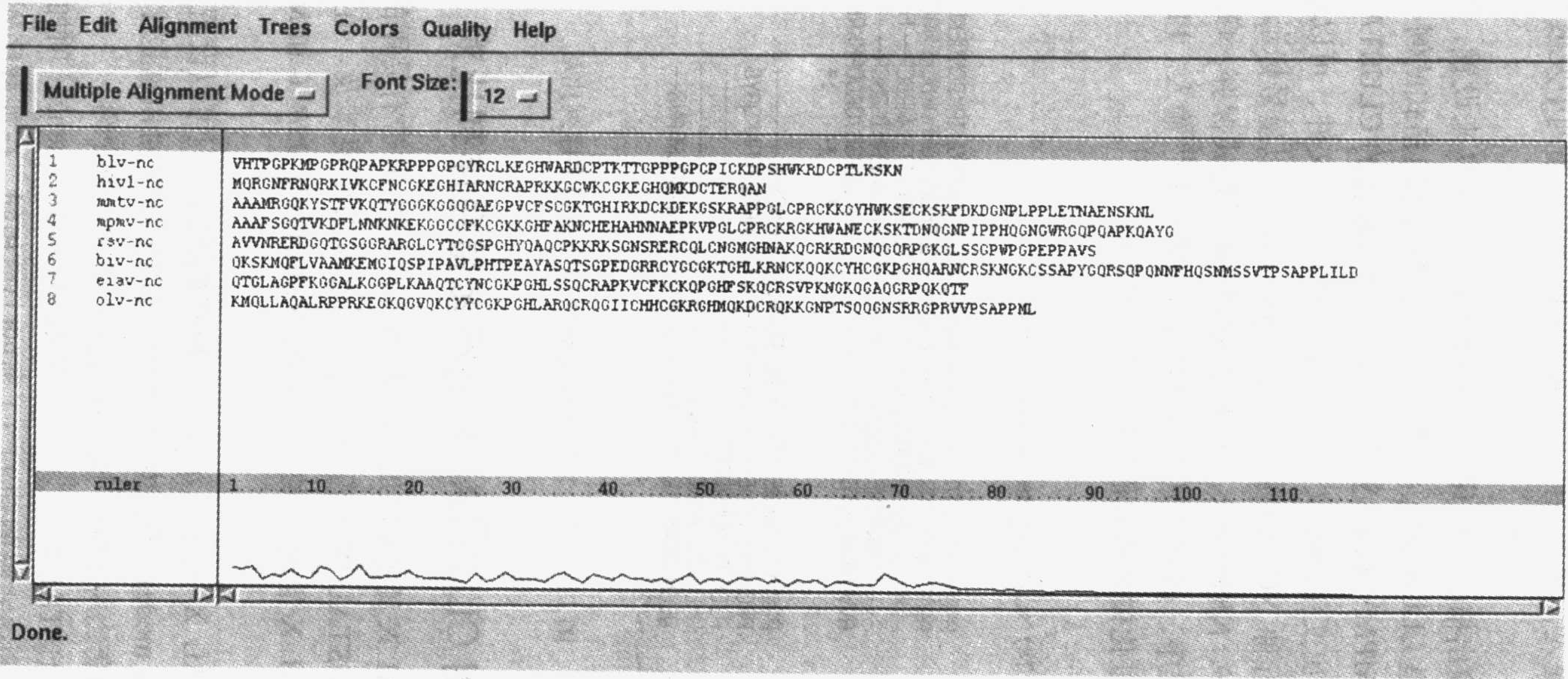


图 11.7　CLUSTAL X 装载序列后准备进行多项比对

序可以通过选择序列名称和 Edit 菜单中的 cut/paste 选项完成。主显示窗口还有标尺以标示残基位置，底部有图表示每一位置的比对质量。该比对图将在下列图 11.9~图 11.11 中详述。

(4) 在 Alignment 菜单下，用户可以进行缺省参数的多序列比对，或利用交互式对话框来修改比对参数，如图 11.3~图 11.5 的 CLUSTAL W 菜单所示。图 11.8 举出了一个定义蛋白空位参数的例子，该对话框中的选项允许用户调节一个特定氨基酸在空位开放位置中的罚分。菜单中的其他选项定义了空位间的最小距离，以及是否将末端空位和内部空位一样对待，以便选择的空位距离分割一致。CLUSTAL W 中和此对话框等效的菜单是在多项比对参数菜单中的选项 8，如图 11.5 所示。

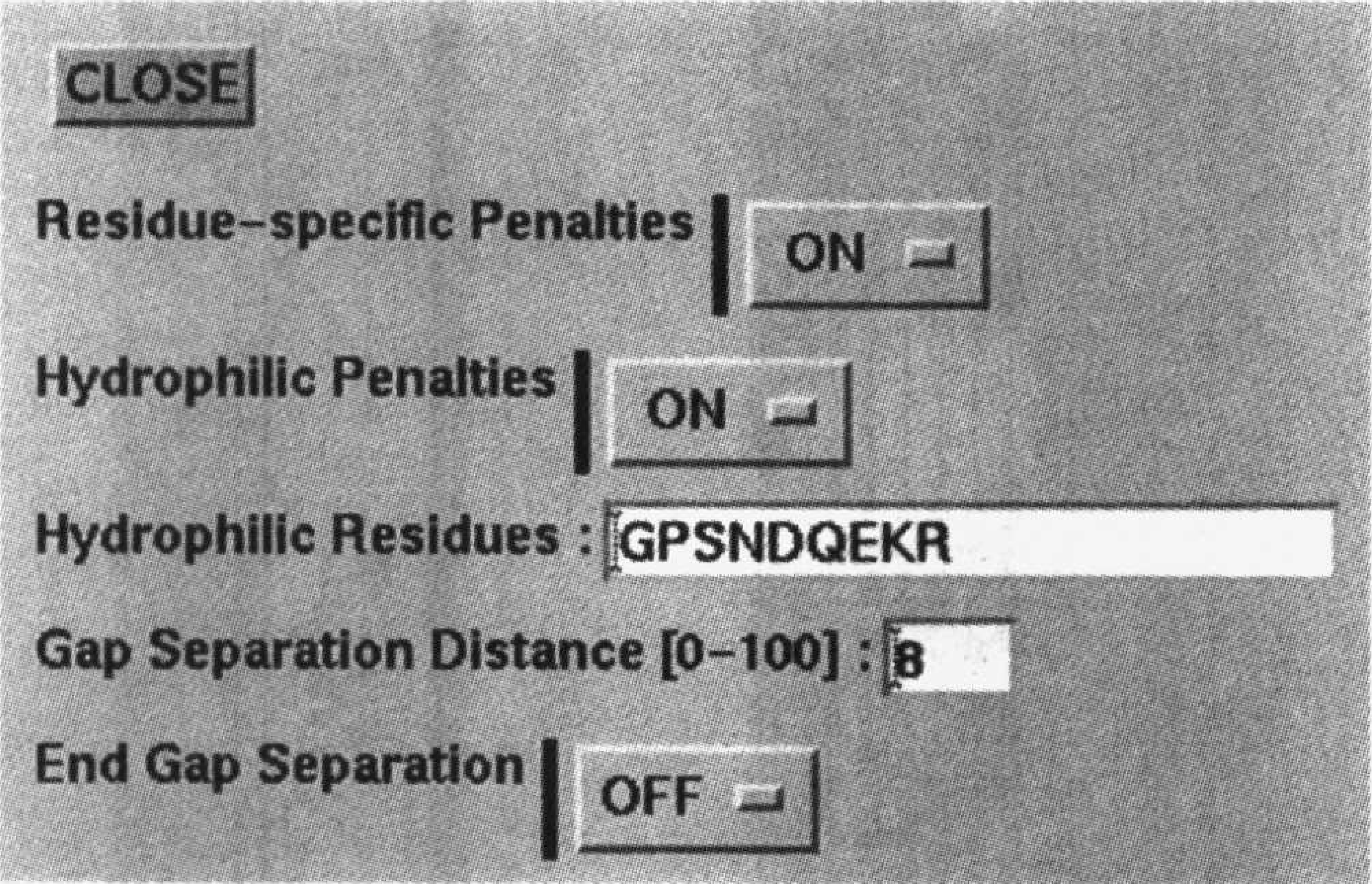


图 11.8　CLUSTAL X 的蛋白质间隙参数

(5) 多重比对后，保守的和比对好的残基在 CLUSTAL X 的主显示窗口中以彩色表示，如图 11.9。颜色是以两类规则来指定的：第一类规则是，一个残基被指定残基特异的颜色和比对中的位置无关；第二类规则是，残基着色是基于每一位置的比对一致序列来定的。该着色系统使比对结果的保守位置以高亮显示。残基着色参数文件在 CLUSTAL X 软件包中提供，但用户自定义文件也可采用。着色参数文件的格式在 CLUSTAL X 的在线帮助中有说明。

(6) CLUSTAL X 具有在多项比对中重新比对错位区域的功能。可以矫正在比对差别很大序列之间比对时无意间引入的错位。这是通过 2 个选项来实现的。已比对的特定序列可以用鼠标单击其名称来选定，如图 11.10 所示。选定序列以白底黑字显示，从多项比对队列中移出，然后重新和仍留在队列中的序列比对。其他选项是让用户从欲重新比对序列中指定一段残基区域。这是通过用鼠标选择想要的残基来实现的，见图 11.10。选好的区域以灰色高亮度显示，从多项比对中移出，用渐进比对方法(如图 11.1)重新比对，然后恰当地插回到队

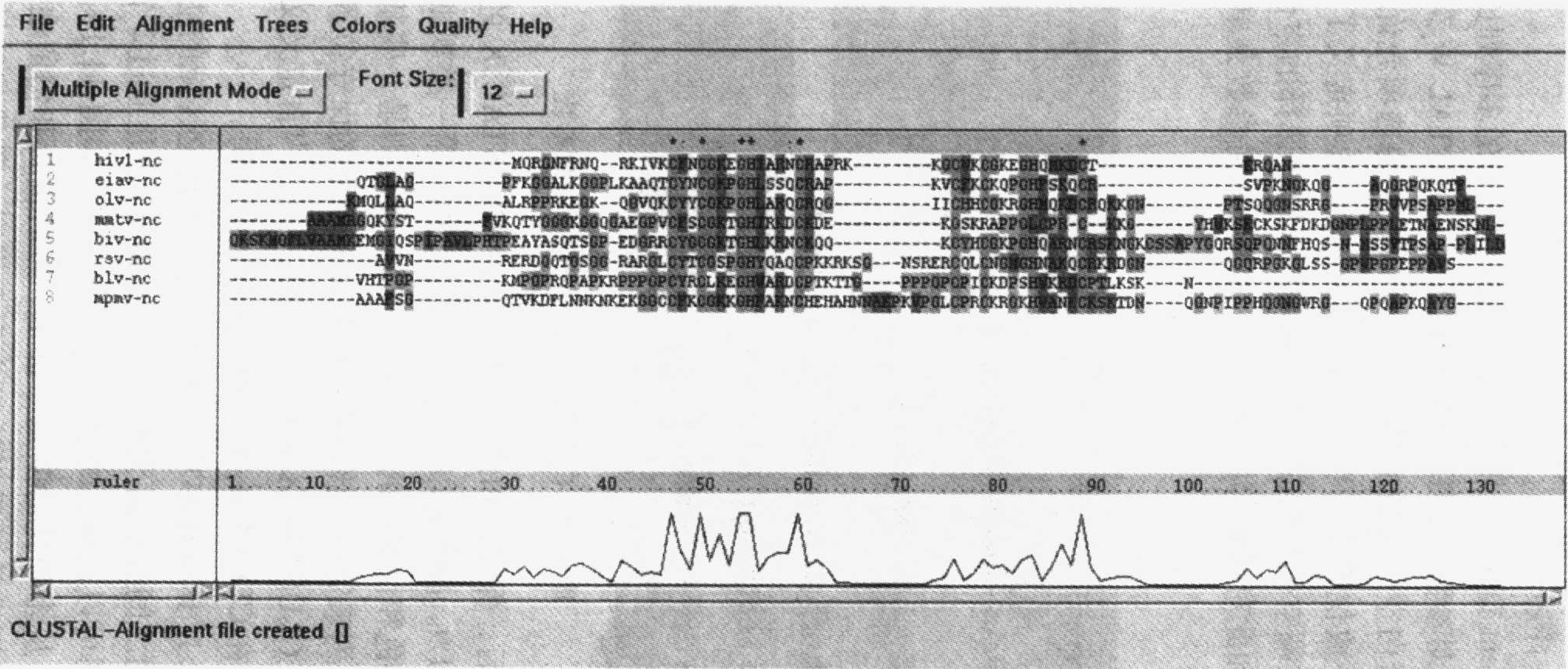


图 11.9　CLUSTAL X 比对窗口显示的比对序列

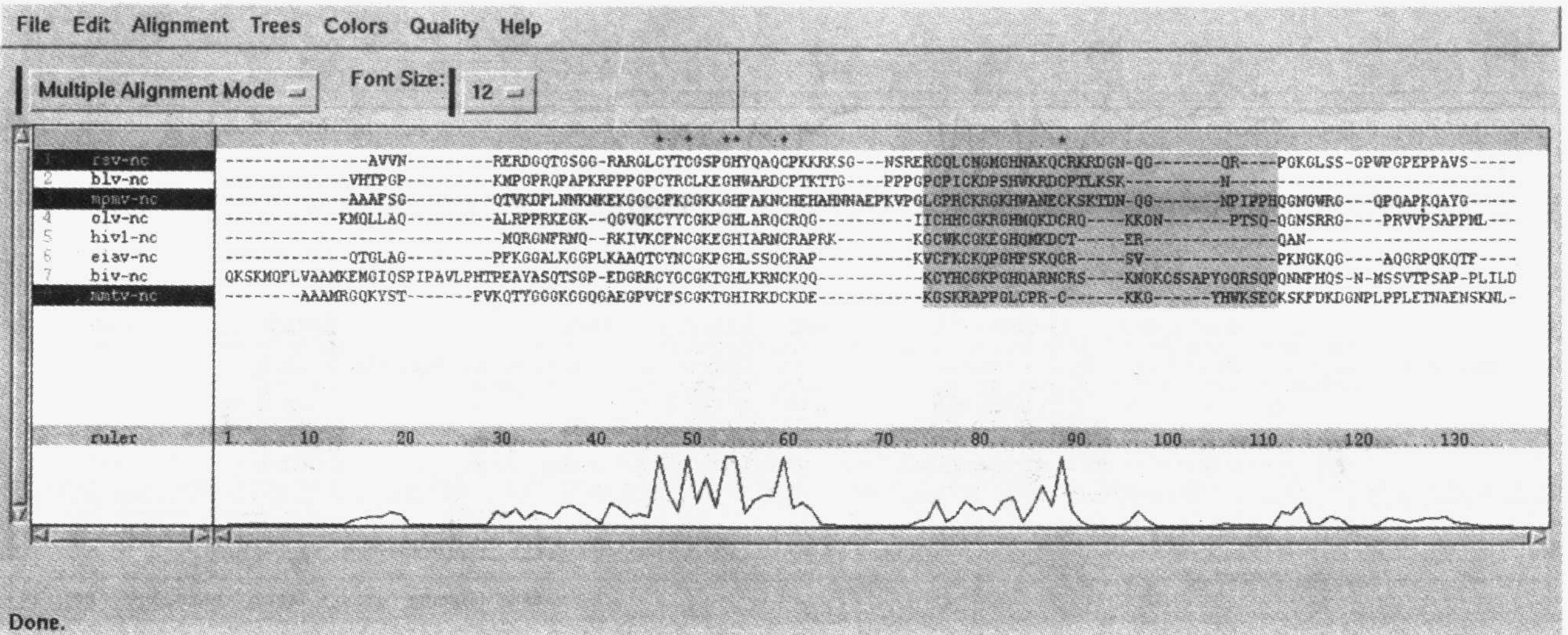


图 11.10　CLUSTAL X 中选择一部分序列进行重新比对

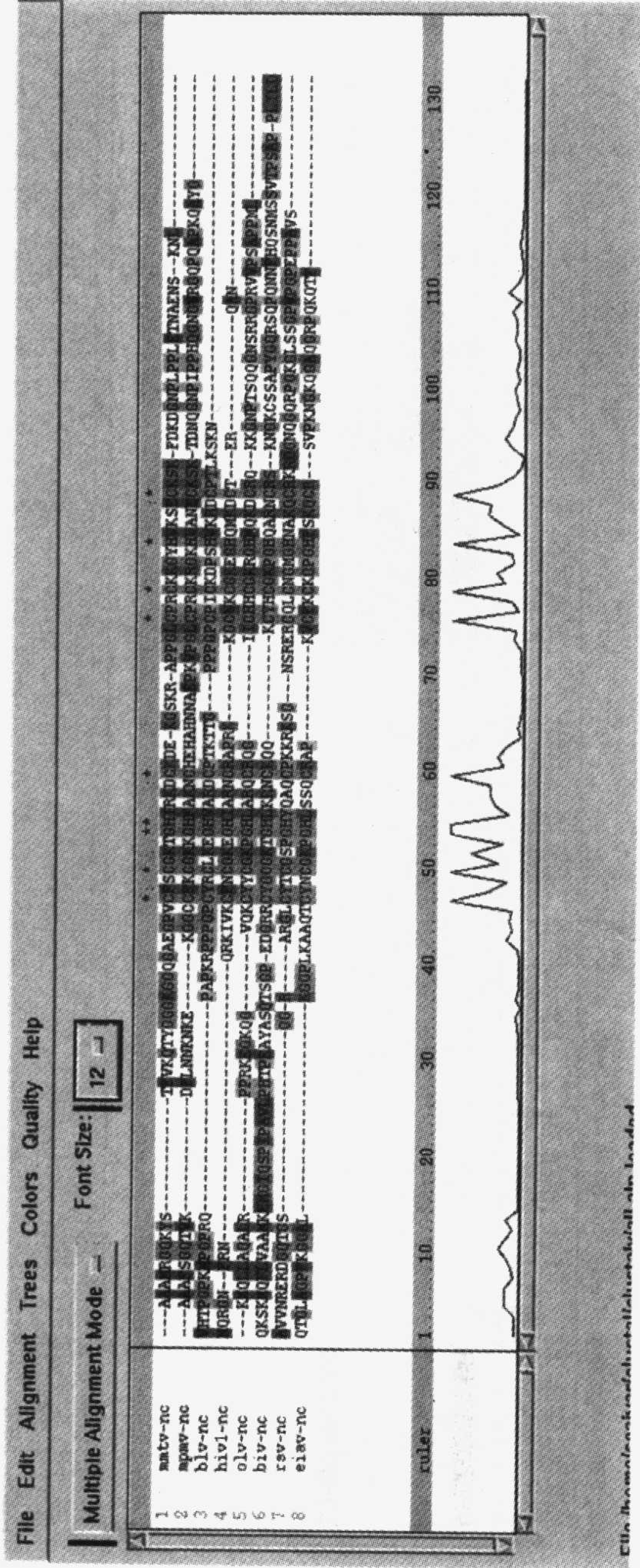


图 11.11　CLUSTAL X 中多项比对可以反复改善

列中。此重新比对选项使原先的比对得以改善和细化。例子见图 11.9 和图 11.11 中 cys-his 盒比对的比较。首先比对所有 8 种 NC 蛋白产生第一次 cys-his 盒比对(并非二次)。重新比对见图 11.10，产生 NC 蛋白第 2 次 cys-his 盒比对，如图 11.11 所示。

(7) 和 CLUSTAL W 不同，CLUSTAL X 可以彩色 PostScript 格式(注：一种用于出版的高清晰度矢量图文格式)输出多项比对结果，用于出版或演示。欲用此功能，在 File 菜单下选择 Write Alignment as PostScript 项。不同的输出参数，如纸张大小、方向、残基颜色、比对输出版面设计等都可以在 PostScript 输出对话框中设定，例子见图 11.12。其他输出格式在 CLUSTAL W[见 11.3.1 节第(7)项] 和 CLUSTAL X 中相同。

11.3.3 利用 CLUSTAL(W/X)进行预先排列文件之间比对

CLUSTAL(W/X)可以用于比对两组现有的排列(profile 预排文件)之间互相比对，或添加新序列到一个现有的排列中。

(1) 为进行预先排列之间比对，在 CLUSTAL X 中切换到相应模式，或者在 CLUSTAL W 的主菜单中选择选项 3(图 11.2)。

(2) 用户可以用 CLUSTAL X 的 File 菜单项下的 Load Profile 1 选项载入原先比对结果文件进行预排文件比对，或者在 CLUSTAL W 的 profile alignment 菜单中选择选项 1，如图 11.13。载入的文件必须是 CLUSTAL(X/W)的多项比对输出文件(*.aln)。第 2 个文件可以是另一个排好的预排列文件或一系列未比对序列。此类文件是用 CLUSTAL X 的 File 菜单中 Load Profile 2 选项，或者用 CLUSTAL W 的 profile alignment 菜单中选择选项 2 载入。

(3) 如果第 2 个文件中含有一个或多个准备和预排文件 1 比对的未比对序列，可以选择 CLUSTAL X 的 Alignment 菜单下 Align Sequences to Profile 1 选项，或者 CLUSTAL W 的 profile alignment 菜单中选择选项 4。

(4) 如果第 2 个文件也是一个预排文件，那么两个文件可以通过在 CLUSTAL X 中 Alignment 菜单的 Align Profile 2 to Profile 1 选项，或者 CLUSTAL W 的 profile alignment 菜单中选择选项 3 进行比对。

(5) CLUSTAL X 的独有功能是序列可以在 Edit 菜单中从一个预排文件中剪切下来，然后再粘贴到另一个文件中。此操作使用户可以从预排文件 2 中选择特异序列插入到文件 1 中进行比对。

(6) 如果有一个确定的结构，就可以用于指导比对。方法是在二级结构选项中提高间隙罚分，这样间隙就会倾向于插入分子表面的环或其他结构未定的区域。

(7) 预排文件比对输出格式和多项比对输出相似，见 11.3.1 节第(7)项。

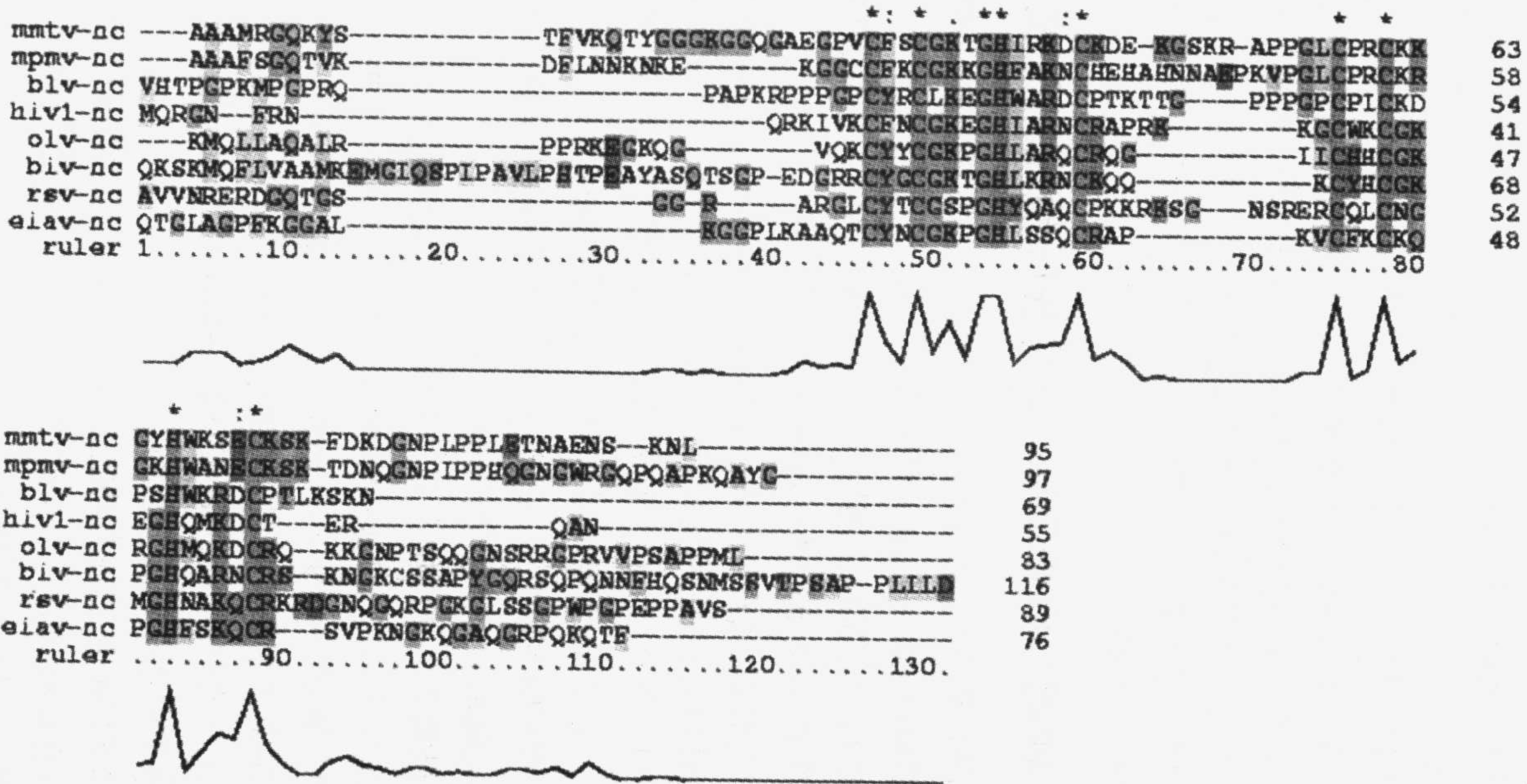


图 11.12 CLUSTAL X 以 PostScript 格式输出比对后序列

```
****** PROFILE AND STRUCTURE ALIGNMENT MENU ******

     1.  Input 1st. profile
     2.  Input 2nd. profile/sequences

     3.  Align 2nd. profile to 1st. profile
     4.  Align sequences to 1st. profile (Slow/Accurate)

     5.  Toggle Slow/Fast pairwise alignments = SLOW

     6.  Pairwise alignment parameters
     7.  Multiple alignment parameters

     8.  Toggle screen display              = ON
     9.  Output format options
     0.  Secondary structure options

     S.  Execute a system command
     H.  HELP
     or press [RETURN] to go back to main menu

Your choice: []
```

图 11.13　CLUSTAL W 的预排比对菜单

11.3.4　系统发生树

CLUSTAL(W/X)可以用原先计算好的多项比对绘制出系统发生树。这是用CLUSTAL X的Tree菜单,或者CLUSTAL W主菜单中选择选项4来进行(图11.2)。绘制树之前必须比对序列。

(1) 用户可以在CLUSTAL X中以File菜单或在CLUSTAL W中选择选项4载入比对序列到内存中，图11.14。

```
****** PHYLOGENETIC TREE MENU ******

     1.  Input an alignment
     2.  Exclude positions with gaps?        = OFF
     3.  Correct for multiple substitutions? = OFF
     4.  Draw tree now
     5.  Bootstrap tree
     6.  Output format options

     S.  Execute a system command
     H.  HELP
     or press [RETURN] to go back to main menu

Your choice: []
```

图 11.14　系统发生树菜单

(2) 欲按缺省参数计算进化树,在CLUSTAL X中的Tree菜单下选择Draw tree now，或者在如图 11.14 中从系统发生树菜单中选择选项 4。其输出文件的扩展名是“.ph”。

(3) 如果要在树计算时去除比对序列都有的间隙位点，以去除比对序列中更加模棱两可的部分,这在比对好的序列差异较大时很有用。此选项在 CLUSTAL X 的 Tree 菜单下，或者 CLUSTAL W 的 phylogenetic tree 菜单的选项 2(图 11.14)。

(4) correct for multiple substitutions 选项可以改正多种替代残基的距离计算。由于序列间有差异，每一个位点都会有多种替代残基，但是只有一种残基是具有代表性的。

(5) 进化树可以解靴带(bootstrap)以给出树中分组的可靠性度量。要求为随机数目发生器输入种子数目和所用的解靴带样本(重复)次数。解靴带的树被写入到扩展名为“.phb”的输出文件中(选择 PHYLIP 输出格式)，或者扩展名为“.njb”的文件(选择 CLUSTAL 输出格式)。

(6) 树可以用 Manolo Gouy 的 NJPLOT 程序(随 CLUSTAL X 程序包发放)显示。NJPLOT 可以输入 CLUSTAL(W/X)输出的 Newick 格式文件“*.ph”或“*.phb”。可以显示树或可以写到 PostScript 格式输出文件中。图 11.12 比对序列的系统发生树输出例子见图 11.15。可以看到慢病毒属(HIV1、OLV、EIAV、BIV)的 NC 蛋白之间的亲缘关系比肿瘤-反病毒的 NC 蛋白近。

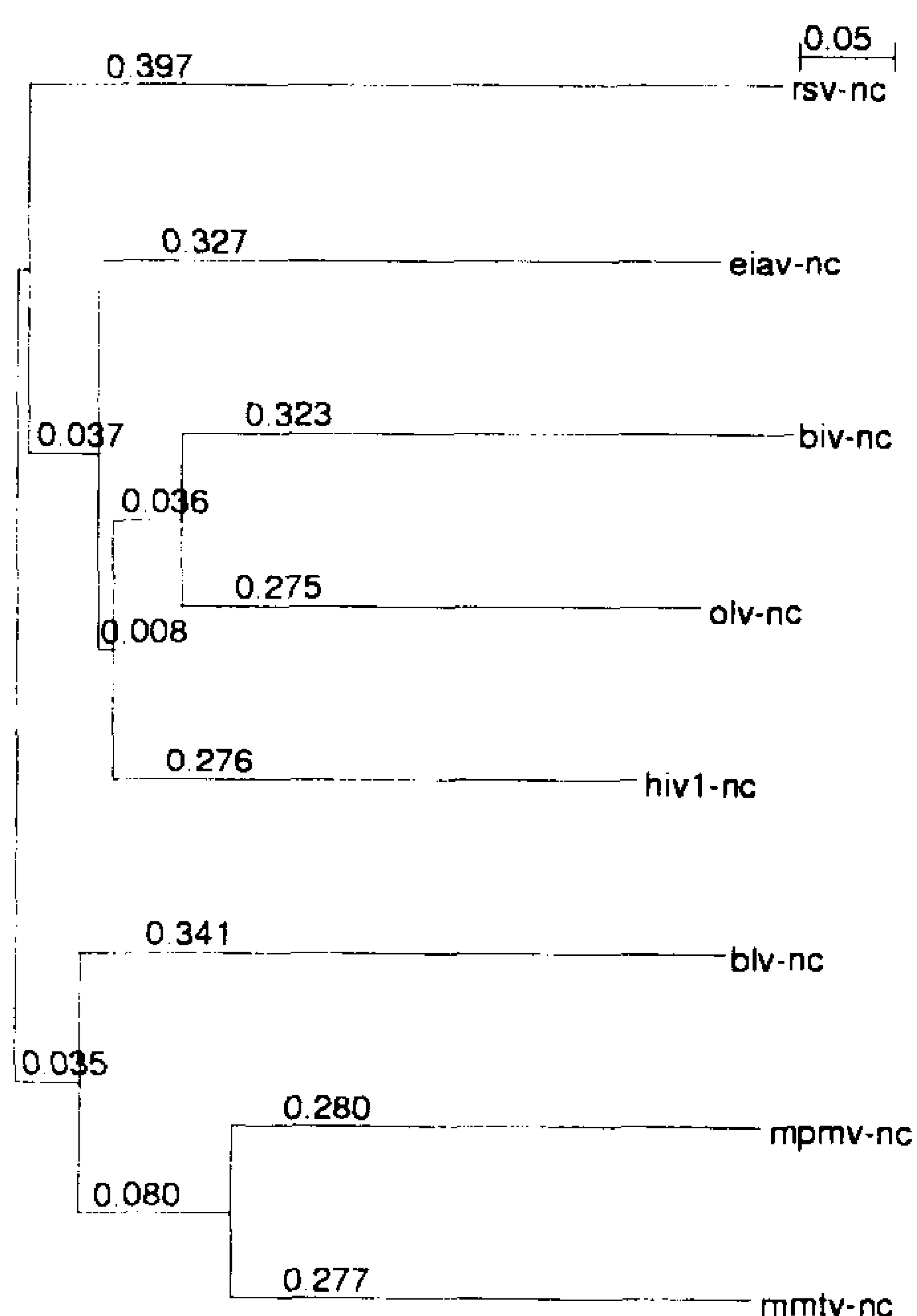


图 11.15　从 CLUSTAL X 多项比对产生的系统发生树

11.4 附注

(1) 读者可参考 Higgins、Thompson 和同仁们写的一系列很好的文章，以进一步了解 CLUSTAL(W/X)程序中所用的多项比对算法[1~4]。

(2) 可以用于序列输入的 7 种序列格式见 CLUSTAL(W/X)的在线帮助文件。除“-”在 CLUSTAL 格式的“.aln”文件中表示一个间隙，“.”在 GCG 格式的 MSF/RSF 文件中表示一个间隙外，所有的非字母/数字符号都会被忽略。读者可以参考“readseq”工具箱，Don Gilbert 编写，可以把许多格式的输入文件转换为 CLUSTAL(W/X)可用的序列。此软件可以从以下 URL 下载：ftp://ftp.bio.indiana.edu/molbio/readseq。

(3) CLUSTAL(W/X)自动检测输入文件中的蛋白质或核酸序列。如果大于 85%的字母都是 A、G、C、T 和 U，则认为是核酸序列；如果蛋白质序列很短，氨基酸组成又很偏(A、G、C、T 多数)，就可能会误认为是核酸序列。

(4) 可以选择 5 种不同的输出格式，如果需要的话，可以同时选定这些格式。CLUSTAL 格式的输出可以在后边作为预排文件比对时读入，或者加入其他序列到排列中。此输出文件由纯 ASCII 文本组成，能被各种文字处理器以文本格式文件读入排版。此格式也和一些序列比对编辑器，如 SeaView[22](ftp://biom3.univ-lyon1.fr/pub/mol_phylogeny/seaview)兼容。要在其他比对编辑器，如 GeneDoc(http://www.cris.com/~ketchup/genedoc.shtml) 中处理多项比对结果，输出比对格式必须是 MSF。

(5) 配对比对可以用慢速/精确(slow/accurate)动态程序算法，或者用快速/大约(fast/approximate)算法[23]进行。慢速/精确算法对短序列很有用，但在多于 30 条以上序列或者大于 1000 个残基以上的长序列处理时会很慢。此时快速/大约算法则很有用。选择此算法时，用户应该调节 k-tuple 和顶-斜线大小。增加 k-tuple 值并且减小 top-diagonal 值会增加速度。

(6) 为进行多项比对，用户可以调节间隙开口和延伸罚分，选择残基权值矩阵。增加间隙缺口和扩展罚分会使间隙更少和更短，但这些取值对末端间隙无效。关于蛋白质，提供了 4 种权值矩阵：identity、BLOSUM、PAM 和 GONNET。GONNET 矩阵事实上是 PAM(Dayhoff)矩阵的现代版本，但是其基于一个更大的数据库。BLOSUM 矩阵是缺省值。对于核苷酸，可以选用 IUB 矩阵或 CLUSTAL W(1.6)矩阵。用 IUB 矩阵时，未知的核苷酸按不确定的 IUB 符号处理(如 RYMWSKDHVBN)。当选 CLUSTAL W(1.6)矩阵时，与不确定的 IUB 符号匹配被视为错配。

(7) 如果有已确定的结构，则可以被用于指导比对，这是通过提高结构元素间隙罚分来完成的。这样间隙就会倾向于插入分子表面的环或其他结构未定的区域。用户提供的间隙罚分设定也可以用于此目的。间隙罚分设定由比对序列的每

一个位点上的一个 1~9 之间的数字组成。对某一给定位置的基本间隙缺口罚分是乘这个数字以获得间隙缺口罚分。间隙罚分设定可以从 CLUSTAL、GDE 和 SwissProt 输入文件。

(8) 系统发生树通过 Saitou 和 Nei[7]的 neighbor-joining(邻接连接)算法计算。CLUSTAL(W/X)可以产生 3 种输出格式的树，但都不是可视化输出，而是 ASCII 文本文件，必须由其他程序读取而绘制出树。CLUSTAL 格式化输出是一个描述格式，列出了所有多项比对序列之间的配对距离和每一配对的比对位置数目。此格式还列出了参与每一次比对步骤的序列和分支长度。PHYLIP 是 New Hampshire 格式，其中列成一系列附件排在文件中，带分支顺序描述，分支长度和序列名称。这些输出文件可以由 NJPLOT 程序(CLUSTAL X 软件包发放)读取。也可以 PHYLIP 软件包中的程序来显示树。这种格式的文件也可以通过其他系统发生树显示程序，如 TREEVIEW AND PHYLO_WIN[22]程序显示。当多项比对结果以 PHYLIP 格式保存后，可以被 PHYLIP 软件包读取产生系统发生树，而且算法不是 neighbor-joining。

(阮承迈　译)

参 考 文 献

[1] Higgins, D. G. and Sharp, P. M. (1988) CLUSTAL: a package for performing multiple sequence alignments on a microcomputer. *Gene* **73,** 237-244.

[2] Higgins, D. G., Bleasby, A. J., and Fuchs, R. (1992) CLUSTAL V: improved software for multiple sequence alignment. *Comput. Applic. Biosci.* (now *Bioinformatics*) **5,** 151-153.

[3] Thompson, J. D., Higgins, D. G., and Gibson, T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight-matrix choice. *Nucleic Acids Res.* **22,** 4673-4680.

[4] Thompson, J. D., Gibson, T. J., Plewniak, F., Jeanmougin, F., and Higgins, D. G. (1997) The CLUSTAL X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res.* **25,** 4876-4882.

[5] Feng, D.-F. and Doolittle, R. F. (1996) Progressive alignment of amino acid sequences and construction of phylogenetic trees from them. *Meth. Enzymol.* **266,** 368-382.

[6] Feng, D.-F. and Doolittle, R. F. (1996) Progressive alignment of amino acid sequences and construction of phylogenetic trees from them. *Meth. Enzymol.* **266,** 368-382.

[7] Saitou, N. and Nei, M. (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4,** 406-425.

[8] Thompson, J. D., Higgins, D. G., and Gibson, T. J. (1994) Improved sensitivity of profile searches through the use of sequence weights and gap excision. *Comput. Applic. Biosci.* (now *Bioinformatics*) **10,** 19-29.

[9] Henikoff, S. and Henikoff, J. G. (1992) Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. USA* **89,** 10 915-10 919.

[10] Dayhoff, M. O., Schwartz, R. M., and Orcutt, B. C. (1978) in *Atlas of Protein Sequence and Structure,* vol. 5, supplement 3 (Dayhoff, M. O., ed.), NBRF, Washington, DC, pp. 345-352.

[11] Benner, S. A., Cohen, M. A., and Gonnet G. H. (1994) Amino acid substitution during functionally constrained divergent evolution of protein sequences. *Protein Eng.***7,** 1323-1332.

[12] Sequences were obtained from *GenBank* with the following accession numbers: 120810 (BIV), 120812 (BLV), 120814 (EIAV), 3023824 (HIV1), 120873 (MMTV), 120876 (MPMV), 120879 (OLV), and 120880 (RSV).

[13] Katz, R. A. and Jentoft J. E. (1989) What is the role of the cys-his motif in retroviral nucleocapsid (NC) proteins? *BioEssays* **11,** 176-181.

[14] Darlix, J. L,. Lapadat-Tapolsky, M., de Rocquigny, H., and Roques, B.P. (1995) First glimpses at structure-function relationships of the nucleocapsid protein of retroviruses. *J. Mol. Biol.* **254,** 523-537.

[15] Bairoch, A. and Boeckmann, B. (1991) The SWISS-PROT protein sequence data bank. *Nucleic Acids Res.* **19,** 2247-2248.

[16] Barker, W. C., George, D. G., Hunt, L. T., and Garavelli, J. S. (1991) The PIR protein sequence database. *Nucleic Acids Res.* **16,** 1869-1871.

[17] Pearson, W. R. and Lipman, D. J. (1988) Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci. USA* **85,** 2444-2448.

[18] Devereux, J., Haeberli, P., and Smithies, O. (1984) A comprehensive set of sequence analysis programs for the VAX. *Nucleic Acids Res.* **12,** 387-395.

[19] Smith, S. Harvard University Genome Center.

[20] Felsenstein, J. (1985) Confidence limits on phylogenies: an approach using the bootstrap. *Evolution* **39,** 783-791.

[21] Felsenstein, J. (1996) Inferring phylogenies from protein sequences by parsimony, distance, and likelihood methods. *Meth. Enzymol.* **266,** 418-427.

[22] Galtier, N., Gouy, M., and Gautier, C. (1996) SEAVIEW and PHYLO_WIN: two graphic tools for sequence alignment and molecular phylogeny. *Comput. Appl. Biosci.* (now *Bioinformatics*) **12,** 543-548.

[23] Wilbur, W. J. and Lipman, D. J. (1983) Rapid similarity searches of nucleic acid and protein data banks. *Proc. Natl. Acad. Sci. USA* **80,** 726-730.

12 用 PHYLIP 进行系统发生学分析

Jacques D. Retief

12.1 引言

系统发生学分析是研究序列之间关系的有力工具。从这些关系中可以推导出基因的起源、进化和可能的结构功能特性上的改变。

PHYLIP(Phylogeny Inference Package)本身是一个集合，广泛地收集了当今进行系统发生分析的几乎所有的方法。此程序包由华盛顿大学(西雅图)遗传学系的 Joseph Felsenstein 分发。可以在许多平台上运行，包括 UNIX 工作站和 IBM 以及 Mac 个人计算机。更有益的是，程序包本身由作者们免费提供。

现在出现的许多系统发生分析程序的数量之多让人目不暇接，而且没有人研究过哪一种方法明显较优。尽管不同方法的准确性可能需要进一步探讨，然而，实际上每一套序列或基因家族都可能有其特殊性，都可能产生最优结果。本章不是作为系统发生学课程，因为已经有许多相关的优秀手册[1~3]，程序带的说明文件也提供了算法和程序选项的详细描述，而是意在对分析特定的基因群找到合适的方法做入门和指导。

12.2 材料

(1) 安装：PHYLIP 软件包已经包括了 UNIX、PC(DOS，Windows 和 Windows NT)的各预编译版本。本章的指导是为 UNIX 写的，但是可以很容易地用到其他平台上。有关硬件支持和下载软件包的详细信息在 `http://www.ibb.waw.pl/docs/PHYLIPdoc/main.html`。运行 DRAWGRAM 和 DRAWTREE 时还需要在本地目录中有一个字体文件，即 fontfile。软件包中已经包括了一系列的字体文件，叫做 font1、font2 等。只要拷贝一个到你的目录里，并且把它重命名为 fontfile 就可以了。

(2) 序列比对：大多数在 PHYLIP 软件包中的程序需要一系列已经比对好的程序作为输入。有许多进行比对序列的软件包在网上可以下载，如 CLUSTAL W(`http://www2.ebi.ac.uk/clustalw/`)[4]。CLUSTAL W 用户

界面带有说明且用法简单。Wisconsin 软件包(遗传学计算机组，[GCG]公司，Madison WI)包括 PIEUP 和 SEQLAB,是一个功能很强的多序列比对程序和编辑器。

(3) 序列重新格式化：READSEQ 程序能把大多数序列改成 PHYLIP 格式。READSEQ 是由 Don Gilbert 在印第安纳大学时设计的，可以通过匿名 ftp(`ftp.bio.indiana.edu/molbio/readseq`)获得。有几个网址还提供了 READSEQ 的网页界面。下列重新格式化指令适用于 GCG 的 MSF 序列格式，因为 GCG 程序是最常用的序列分析包之一。此程序是普遍适用的，可以很容易地改成为其他格式使用。

(4) 图形输出程序：许多商业程序，如 Macromedia FreeHand®和 Corel Draw®能输入和编辑 PHYLIP 生成的图形。

12.3 方法

12.3.1 选择合适的序列

(1) 所有系统发生学分析的一个基本假设，就是比较的基因必须是定向同源进化基因(orthologous gene)。这是很明显的，例如，易于水平转换的基因或孤儿基因(orphan gene)，将产生伪性结果，因为它们从祖先基因进化来的约束条件不同。因此，首先确认要比较的基因归于同一类。

(2) 当缺乏祖先序列时外围组(outgroup)提供了用于测量距离并帮助找出树根的参考。所谓外围组就是最相近但不属于所研究的组。例如，为构建一个哺乳动物序列树，一个鸟类的序列可以提供适当的外围组，而植物序列就不适合，因为植物亲缘关系太远，会使比对出错，降低距离估计准确性。

(3) DNA 还是蛋白质？当太多突变发生时序列之间差异就会太大。例如，在一个位点从状态 A 突变到状态 B，当更多突变发生时，突变成状态 C 或回复成状态 A 的概率就会增加，使我们对突变的数目估计不足。除去遗传密码的冗余，蛋白质通常是基因的有功能产物，保留蛋白质功能是序列保守的原动力。因此蛋白质序列的改变要比 DNA 序列慢得多，是研究远缘序列或变化得特别快基因的首选。在有些情况下，当一个基因改变得非常慢时，或者当关系非常近的序列时，或者要研究的基因非常小时，多肽序列就可能包含了太少的信息，不能用来解析进化树，此时 DNA 序列就是较佳的选择。

12.3.2 序列比对

输入序列的比对是进化树分析的基础。比对的错误可能使最精心设计的算法失效。不幸的是，序列比对程序总是不完善，获得满意的比对可能要花去分析过程的多半时间，尤其当比对结果有许多间隙时，这样的时间分配就是合理的。注

意下列常见的序列比对问题：

(1) 序列特征，如起始密码子、终止子和内含子接头应当特别注意。PILEUP和CLUSTAL程序不考虑ATG序列作为起始密码子或编码甲硫氨酸或是读框外序列的相对重要性。结果序列特征，如起始密码子不会完整地比对。当起始密码子位置移动时，在进化上的含义是很深刻的，因为可能引起框移或缺失，从而改变基因产物的大小。这些过程发生的机会远远小于相邻间隙被错误比对。观察被错误比对的起始和终止密码子以及内含子接头周围的序列，如果考虑到其重要性就应该完全改善为止。蛋白质序列的已知功能域也是如此。

(2) 间隙的处理也是一个问题。间隙，当以“-”编码时，其他程序会认为是附加的字符。例如，DNAPARS 识别 5 个字符 A、T、G、C 和间隙。当 2 个序列都共有一段比如 50 个核苷酸的间隙时，程序会认为那段 50bp 序列是完美的配对，然而事实上间隙可能是由一个简单的过程生成的。如果没有补偿的话，共有大段间隙的序列会归到同一组中。“？”号用于编码丢失的序列，如当长度不同的序列之间比较时序列的末端。如果仅仅一个“-”字符被“？”取代，此间隙将仅被算成是单一事件，降低间隙对系统发生树的影响。如果序列含有许多间隙，把要比较的两条序列间隙分别标记为“-”和“？”是很好的方法，以正确地得出间隙对进化树的影响结果。间隙本身不应该决定最后生成的进化树。

(3) 编码区域的比对应该用蛋白质区域进行比较。蛋白质区域的比对应该和核苷酸编码区的比对重新协调。现在还没有程序能自动做这件事，所以需要用序列或文本编辑程序手工编辑。DNA 比对时，这将保证间隙是以三联体形式出现。蛋白质序列中，密码子使用可能有助于解决模棱两可的比对(图 12.1A 和图 12.1B)。

(4) 低信息区域，如内含子可能含有大片的双核苷酸或单核苷酸。这些片段可能被很快地插入或延伸，因此要仔细对待。类似地，大量的相等有效的比对区域可能影响结果(图 12.1C)。低信息区域从比对序列中删除后效果会很好，因为去除了它们产生的随机偏性。但删除部分比对序列会影响到估计的分支长度。

A

```
T  Y  R  R  S  R
T  Y  R  R  S  R
T  Y  R  -  S  R
T  Y  R  -  S  R
T  Y  R  R  S  R
```

B

```
ACA TAC AGG CGA AGC CGG
 T   Y   R   R   S   R
ACA TAC AGG CGA AGC CGG
 T   Y   R   R   S   R
ACA TAC AGG --- AGC CGG
 T   Y   R   -   S   R
ACA TAC --- CGA AGC CGG
 T   Y   -   R   S   R
ACA TAC AGG CGA AGC CGG
 T   Y   R   R   S   R
```

C

```
gatttggggtggg
gattattggggaag
aattat---ggtgg
gatctagtttatgg
gatt-tggggtgg
g--tattggggtgg
gatttggggaggg
```

图 12.1　比较比对编码区和比对的蛋白质序列可解析模棱两可的间隙

A. 第 4 列的间隙排列显然是人为的，排在第 3 列也相同；B. 蛋白质序列和核苷酸序列比较解析了比对并且显示了间隙被蛋白质序列的第 3 列和第 4 列间分开，导致可能影响吝啬树(parsimony tree)结构的有意义位点丢失；C. 一个域中有大量的同等有效比对存在时，这样的比对产生随机噪声，删除这样的域可以改善分析结果

12.3.3 格式化序列

PHYLIP 软件包采用自己独有的序列格式(图 12.2)。此格式相对较简单，你可以重新格式化一系列比对好的序列成 PHYLIP 格式，所需要的仅仅是一个文本编辑器和你的判断。但是，还有更简单的选择，就是选用重新格式化程序，如 READSEQ。READSEQ 接受大多数共同序列格式，但 READSEQ 并不完善，并且序列需要在转换之前进行某些准备工作。下列步骤将把一个按 GCG 的 MSF 格式比对的多序列 DNA 或蛋白质序列转换成 PHYLIP 格式。大多数步骤将也可以用于其他序列格式。

```
6 50
HSP1_PANTR    ARYRCCRSQS RSRCYRQRQR SRRRKRQSCQ
HSP1_GORGO    ARYRCCRSQS RSRCYRQRQT SRRRRRRSCQ
HSP1_HYLLA    ARYRCCRSQS RSRCYRRGQR SRRRRRRSCQ
HSP1_HUMAN    ARYRCCRSQS RSRYYRQRQR SRRRRRRSCQ
HSP1_RABIT    ?????CRSQS RSRCRRRRRR CRRRRRRCCQ
HSP1_SAGIM    ARYRCCRSQS RSRCYRQRRR GRRRRRRTCR

TQRRAMRCCR RRSRMRRRRH
TQRRAMRCCR RRNRLR????
TRRRAMRCCR PRYRLRR???
TRRRAMRCCR PRYRPRCRRH
-RRRVRKCCR RTYTLRCRR?
-RRRASRCCR RRYKLTCRR?
```

图 12.2 PHYLIP 使用的序列格式

是一个典型的分栏输入序列。比对序列的顶部表示序列的数目和序列中字符的个数，字符都用大写。序列末端的间隙用丢失数据“?”字符表示，而序列内部的间隙用“_”号表示，空格被忽略或者被包含以改善序列的可读性

(1) 把 MSF 文件装载入文本编辑器，例如 JOVE、EMACS 或 VI，特点是必须能自动寻找和替换字符。在下列步骤中要注意不能改文件头，序列和文件头之间用一行由“//”组成的行分隔。仅修改这行下边的字符，并且修改完毕后保存文件。

a. 把所有的小写序列字符用大写的替换。如果用 GCG 程序更方便，只要在比对之前用 ONECASE 命令(UNIX：`onecase -men=ufilename`)。

b. 用“?”或“-”符号替换所有的“~”，见 12.3.2 节第(2)项。

c. 用“-”符号替换所有的“.”符号，记住千万不要替换 GCG 命令行的“..”号。

d. 确认序列中没有模棱两可的或未知符号。例如，DNA 序列中只能有 A、T、G、C 和间隙符号，空格和数字会被忽略。

(2) 改名称为 filename. msf 的 msf 格式时按如下操作。

a. 键入：readseq filename.msf。

b. 当提示时键入输出文件名，如 filename. inf。

c. 选择第 12 项 PHYLIP 为输出格式。

d. 当提示时确认所有的序列名称都列出并键入“all”。

(3) 如果文件被不适当地格式化，你就会收到 PHYLIP 程序神秘的“Memory allocation error(内存分配错误)”信息。并非所有的 PHYLIP 模块都对文件格式严谨。也可能用 SEQBOOT 能成功转换但用 PROTPARS 时则会出错。此时，把序列载入文本编辑器，或在 UNIX 中键入：more filename.inf。按下列步骤核对文件。

a. 如果使用字处理器，如 Microsoft Word，确认文件以纯文本文件保存。

b. 如果比对在改格式过程中被弄乱了，你可能是忘了替换“~”符号。

c. 确认使用了正确的文件类型，并且所有的程序需要序列文件，如 FITCH，是用含有 DNADIST 或 PROTDIST 生成的距离矩阵的文件。流程图表示了输入文件要求。

d. 所有序列符号必须大写。

e. 序列中的间隙必须用“-”或“?”表示(“.”是不允许的)。

f. 序列中只能含有合法字符，否则会提示“Illegal character”错误。

g. 第一行包含序列数目，后跟序列中的字符数。

h. 所有序列必须等长，不够的地方用“-”或“?”补齐。

i. 序列名称必须是正好 10 个字符长，不够则用空格补齐。

j. 如果还有错误，则用 READSEQ 再过一遍，有时会改正错误。

12.3.4 统计学方法

没有统计学方法，就不可能判断特定进化树上的分支点(图 12.3A)。Felsenstein 介绍了系统发生分析的解靴带(bootstrapping)方法[5]。这是一个把原数据包重新采样，生成一系列数据包，每一包中一些数据被随机改动。这和删除不同，改动保证数据包大小不变。在 PHYLIP 3.6 版本中设置选项 b 为 3，保证每一密码子位置都以相同的频率采样。在最简单的形式中，可以想像成一个基于数个易受扰动的位点细致的分组。而当某些位点被破坏时，联系仍然维持。通常 SEQBOOT 程序会生成 100 个数据包。把 m 选项设置为 100 即可生成 100 个树，从这 100 个树中再用 CONSENSE 程序生成一致序列的树。输出文件的解靴带值表示 100 个树中一个特定分支点上的发生数。如果我们在图 12.3B 中插入解靴带值，就可以看清哪一个分支点是有效的。解靴带值高于 90(每 100 个数据包)则认为有统计学意义，低于 50 则基本上是随机的。在图 12.3 中，啮齿类是唯一的有意义分组。一个更忠实树的形式是通过把分支糅合为一个 polytomy(图 12.3C)，这可以通过去除树文件中的圆括号来实现。RETREE(在 PHYLIP 3.6 版本中独有)使你很容易地去除无意义的分支点。

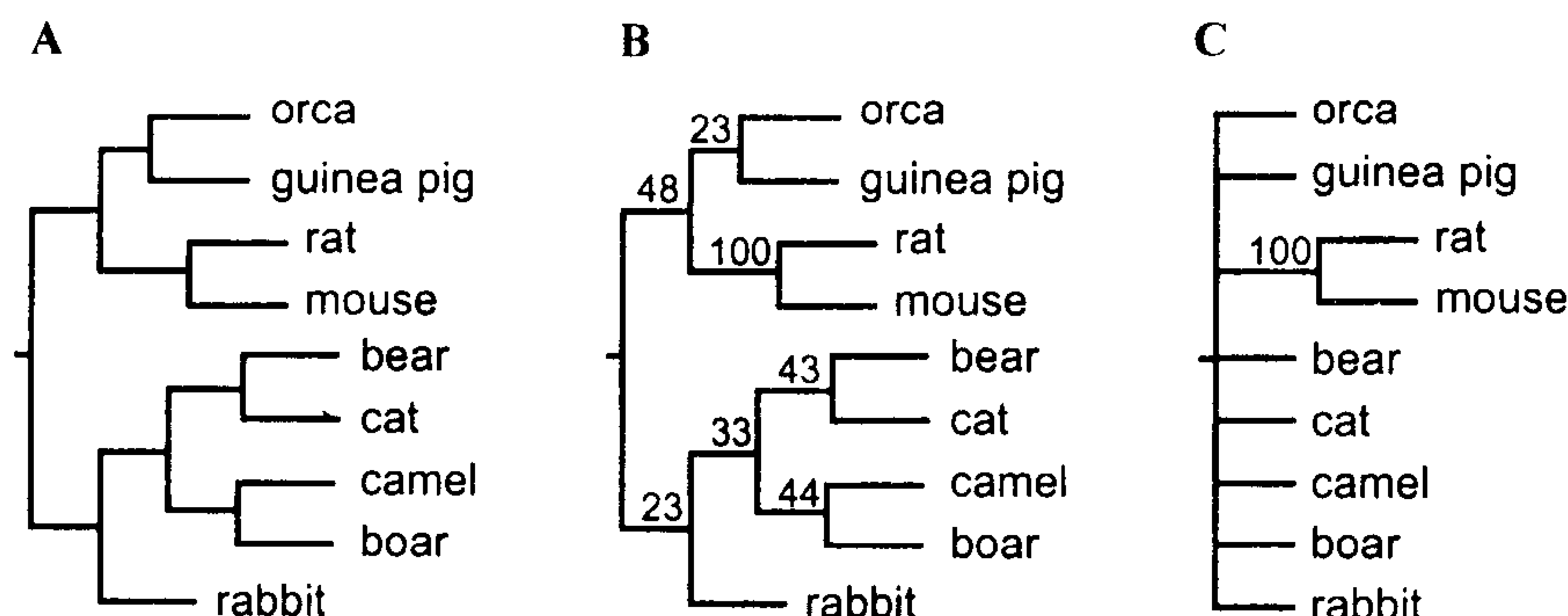


图 12.3 解靴带值表示在一个随机重新采样的数据包中某一分支出现的次数

A. 此蛋白质吝啬树(parsimony tree)显示许多很不平常的联系，如把杀手鲸和啮齿类分在一组。如果没有统计学方法，就不能判断树的有效性。B. 相同的树节点加上了从 100 个数据包中得到的解靴带值。(解靴带值是从 CONSENSE 程序的输出文件转换过来的。)现在很清楚，除了大鼠和小鼠，其他所有的联系都是随机的。C. 相同的树上把所有值小于 50 的分支都去除成为一个多歧树，这是一个真实的树模型

12.3.5 运行程序

执行 UNIX 程序时，先键入程序名然后回车。在图形用户界面，如 Windows 中双击程序图标即可。程序会从一个文件中载入输入，而在另一个文件中输出。通常这些文件称为输入文件(infile)、树文件(treefile)或输出文件(outfile)。当程序执行时会出现一个选项菜单。改变选项时，键入选项符号然后回车。需要的话，程序会提示输入。当所有的设置都无误时输入“y”然后回车。程序的数目之多和大量的选项让人望而生畏。图 12.4~图 12.6 中的流程图是一个对输入文件和输出文件要求的指导，通常程序将提供较佳的缺省值作为起点。

下面所有的命令都是 PHYLIP 3.5c 版本的，因为此时 PHYLIP 3.6 版本还是一个“pre alpha”版本(测试版本：译者注)。PHYLIP 3.6 版在功能上和原先版本相同，只是加上了一些新的特点和选项。重要的改进在附带的文本中有说明。一个有用的改进是更改输出文件的选项。重要的是，树文件现在一贯地称为 intree 和 outtree。这些改变在流程图中有说明，其中版本 3.6 中的文件名是用白字黑底表示。要想知道现在运行的是哪一个版本，只需看看菜单顶部的版本号。

12.3.6 吝啬算法(parsimony method)

此算法是一种基于字符的分析，每一字符是和邻近字符无关的，并且只有信息位点才予以考虑。对于有信息位点同样的突变必须出现在至少 2 个序列中。有关吝啬算法详细的描述见参考文献[1]、[3]、[6]。吝啬算法程序计算树分支的次序，并不给出分支长度。吝啬算法的优点是采用了逻辑模型并且计算快速；缺点是大量的数据被丢弃，因为只有信息位点才予以考虑。这样如果用的是短序列或者信

息位点太少的话就可能出现问题。图 12.4 显示的是吝啬算法的流程图。下边列出了典型吝啬算法的步骤：

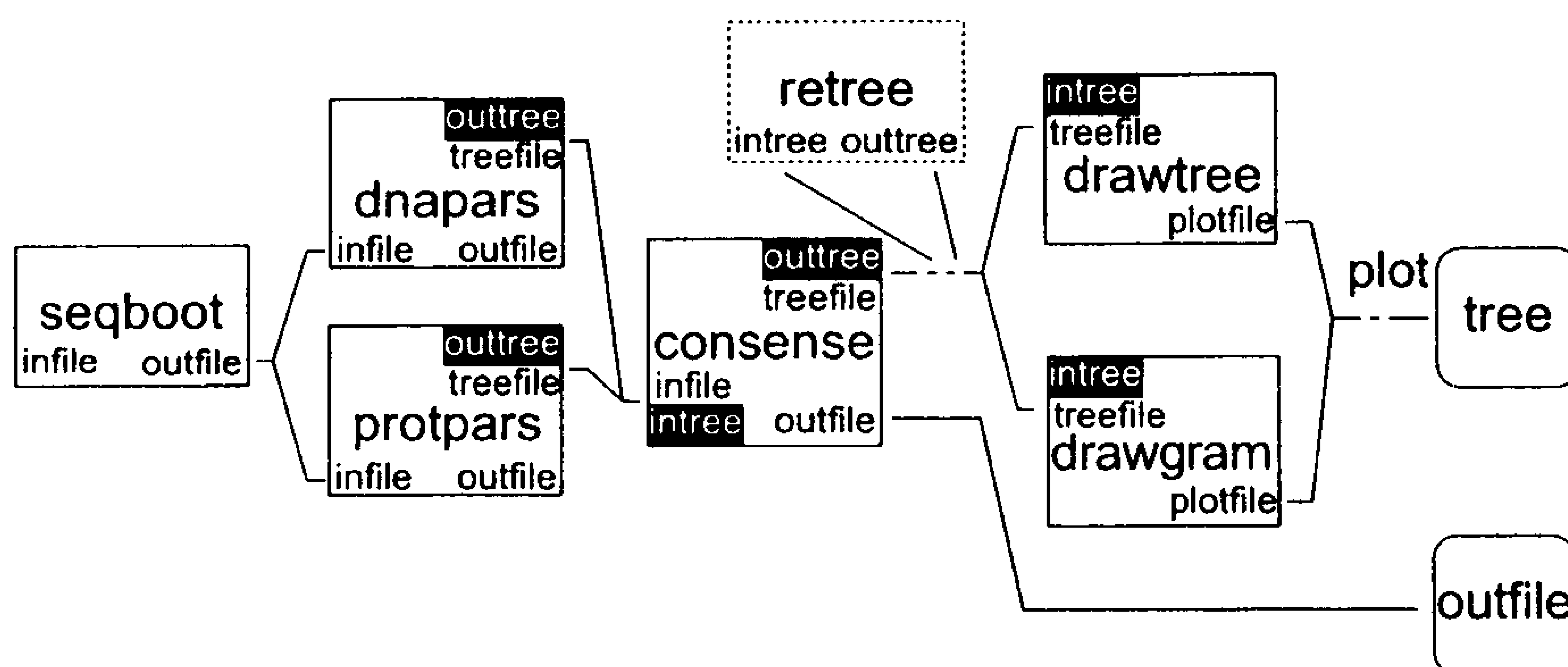


图 12.4 表示程序之间的组合以生成一个吝啬树的流程图

此图是这样读的：SEQBOOT 程序从 infile 文件中获得输入，并把输出写到 outfile 文件中。第 2 个模块是用来匹配序列文件类型的。进行第 2 步，即 DNAPARS 时，SEQBOOT 的 outfile 应该拷贝成 infile，因为 DNAPARS 也从 infile 中获得输入。反过来，DNAPARS 会生成一个 outfile 和一个 treefile。SEQBOOT 和 CONSENSE 都可以选择用于解靴带的程序。RETREE 也是一个可选的用于控制树的程序；DNAPARS 和 PROTPARS 分别用于 DNA 或蛋白质序列；DRAWTREE 和 DRAWGRAM 是按照想要树的形式而选用的。最后的 outfile 中有解靴带值。黑底白字表示 PHYLIP 3.6 版本所需要和生成的文件名

(1) 把 PHYLIP 格式的序列比对文件拷贝到名为 infile 的文件中(UNIX：`cp yourfile infile`)。文件格式见 12.3.3 节。

(2) 可选地，如果要采用解靴带算法数据包运行 SEQBOOT(UNIX：`seqboot`)。提供一个随机数目并接受缺省值生成 100 个数据包(见 12.3.4 节)，把 SEQBOOT 的 outfile 文件拷贝到 infile(UNIX：`cp outfile infile`)。

(3) 运行 DNAPARS 或 PROTPARS 进行吝啬计算，计算出一个 DNA 或蛋白质数据包(计算 DNA 时 UNIX：`dnapars`，计算蛋白质时 UNIX：`protpars`)。弹出含数个选项的窗口：

a. 选项 U 可以指定自己的树，文件格式参考 PHYLIP 说明文件，可以计算吝啬树然后采用距离树的分支长度，最后接受缺省值让程序找出最佳树。

b. 选项 J 允许序列次序随机化。树是自顶向下构建的，密切相关的序列应该放在比对的顶部，这样可以先被分组；关系较远的序列放在比对的底部，这样被最后加到树上，干扰其他分组的机会减低。GCG 的 PILEUP 程序比对好的序列是自动按此模式安排的。随机化序列次序会消除构建树过程中的序列次序偏差。

c. 选项 O 可以指定一个输出组，输出组的选择细节见 12.3.1 节第(2)项。

d. 选项 T 指定吝啬阈值，可以限制计算远缘分支的步骤数。

e. 选项 M 指定用于解靴带算法的序列套数，如果使用解靴带数据记得把它设置成 100。

f. 选项 I 指定缺省序列格式，具体见图 12.1。

g. 选项 0~6 指定不影响数据的系统参数，缺省值在大多数情况下都适用。

(4) 如果选择使用解靴带数据，则把树文件拷贝成 infile (UNIX：`cp treefile infile`) (PHYLIP 3.6：拷贝树文件成 intree，UNIX：`cp outtree intree`)。然后运行 CONSENSE(UNIX：`consense`)。CONSENSE 程序生成的输出文件即包含了解靴带值。

(5) 用 DRAWGRAM 或 DRAWTREE 绘制树。这些程序需要由 CONSENSE 生成的单个树文件(PHYLIP 3.6：拷贝树文件成 intree，UNIX：`cp outtree to intree`)。如 DRAWGRAM(UINX：drawgram)，结果输出到 plotfile 文件中。(PHYLIP 3.6：下列 3 个菜单被结合到一个中，但是选项基本相同。)

a. 第一个菜单可以选择绘制树的设备，应该选用 PostScript 打印机。

b. 第二个菜单选择预览树的设备，如果你的显示器只能显示字符，选 N。

c. 第三个菜单确定绘制树的选项，绘制 phenogram 时，先选树形式(tree style)，然后选 P。

d. 其他选项按自己的喜好改变。

(6) 用于绘制图形文件的算法根据操作系统和 DRAWGRAM 或 DRAWTREE 程序的选项来确定。如果有 PostScript 打印机并且选择 LaserWriter，你仅需把 plotfile 文件拷贝到打印机(UNIX：`lpr-Pprintername plotfile`)。

12.3.7 最大似然算法

计算最大相似性的基本步骤和吝啬算法相同。每一位点都予以考虑，并且从核苷酸池中替换特定核苷酸的相似性都被计算[7, 8]。算法也适应于蛋白质序列。其优点是每一位点都予以考虑。甚至不变化的位点都有机会改动并且改回原来状态。DNAMLK 和 DNAML 相同但采用了一个分子钟。还有高级选项决定特定树的似然性。在分析中所有位点都很重要，即使是不变化的位点，以给出分支长度的准确估计。缺点是算起来很慢。典型的 DNA 最大相似性算法的步骤如下(图 12.5)。

(1) 拷贝 PHYLIP 格式的比对序列到 infile 文件中，此例中序列比对文件名称为 yourfile(UNIX：`cp yourfile infile`)。见 12.3.3 节的文件格式说明。

(2) 可选地，如果要采用解靴带算法数据包运行 SEQBOOT(UNIX：`seqboot`)。提供一个随机数目并接受缺省值生成 100 个数据包(见 12.3.4 节)。把 SEQBOOT 的 outfile 文件拷贝到 infile(UNIX：`cp outfile infile`)。

(3) 运行 DNAML(算 DNA 时 UNIX：`dnaml`)。弹出含数个选项的窗口。

a. 选项 U 可以指定自己的树，标准分析接受缺省值让程序找出最佳树。

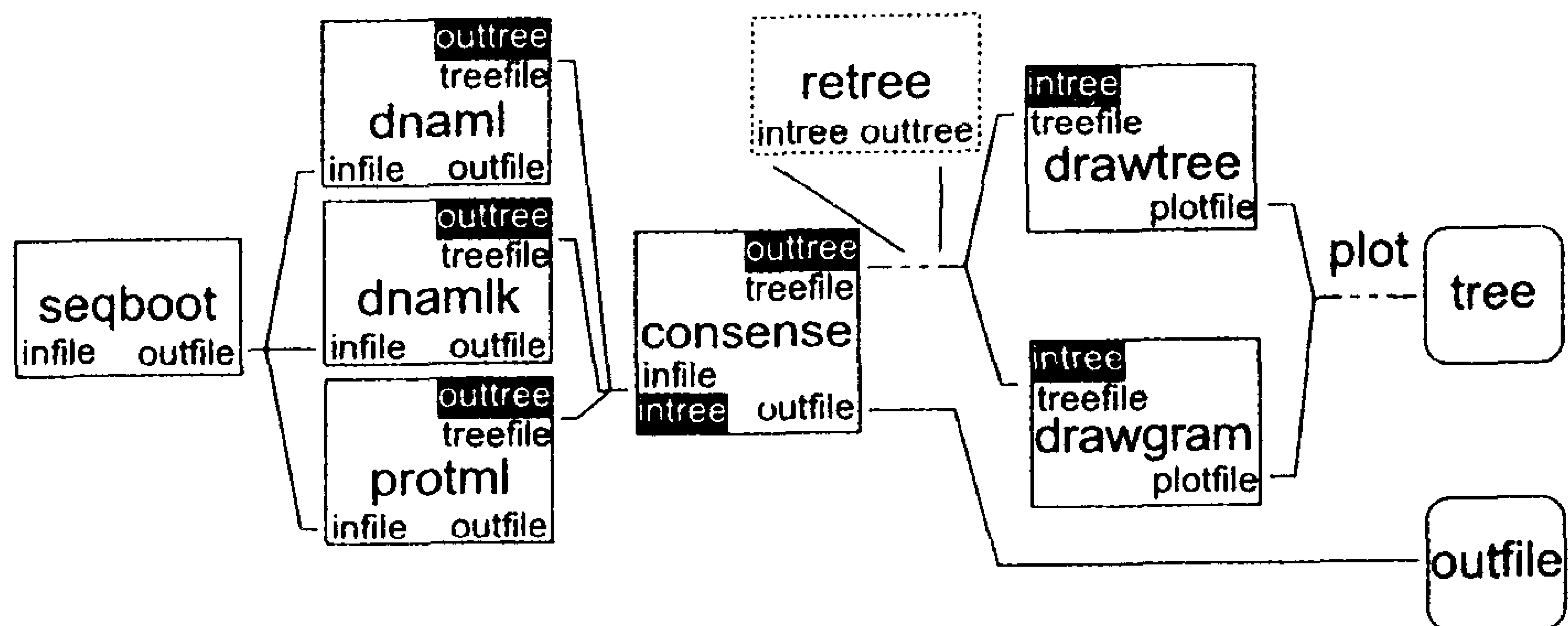


图 12.5 生成最大相似性树的流程图

此法和图 12.4 的吝啬树算法很相似

b. 选项 T 设置转换/颠换比率，缺省值 2 就很好。

c. 选项 F 设置自己的碱基频率，频率必须加起来为 1 并且输入到一空格隔离的行中。缺省的经验频率是从输入序列中计算而得，尽管不是一个真的最大相似性值，但也通常很接近。

d. 选项 C (PHYLIP 3.6 中的 R 和 W)，允许用户设置范围和比率的数目，是为高级用户准备的，用户必须了解它们的含义。细节见 PHYLIP 说明文件。

e. 选项 S (PHYLIP 3.6)，提供了对最佳树一个快速的，但较不严谨的估计。实际上，对于大多数分析来说都适用。

f. 选项 G，去除并随后加上树上的每一分组，确认每一分支都重新计算并且位置优化。选项 O 可以指定一个输出组，输出组的选择细节见 12.3.1 节第(2)项。

g. 选项 M 指定用于解靴带算法的序列套数，如果使用解靴带数据记得把它设置成 100。

h. 选项 J，O，I 和 1~4 是常用的系统选项，在其他地方讨论。缺省值在大多数情况下都适用。

(4) 如果准备使用解靴带数据,则把树文件拷贝成 infile(UNIX：`cp treefile infile`)(PHYLIP 3.6：拷贝树文件成 intree，UNIX：`cp outtree to intree`)。然后运行 CONSENSE(UNIX：`consense`)。用选项 O 或 R 确定树根或者采用输出组(outgroup)。

(5) CONSENSE 程序生成的输出文件即包含了解靴带值，树文件写入到 treefile 中，在 PHYLIP 3.6 版本中树文件写入到 outtree 文件中。

(6) 用 DRAWGRAM 或 DRAWTREE 绘制树。这些程序需要由 DNAML 或 DNAMLK 程序以及采用解靴带数据时的 CONSENSE 程序生成的单个树文件。见 12.3.6 节第(5)项的例子。DRAWGRAM(UINX/DOS：`drawgram`)。注意 PHYLIP 3.6 中拷贝树文件成 intree，因为这是它的输入文件(UNIX：`cp outtree intree`)。

(7) DRAWGRAM 或 DRAWTREE 的结果输出到 plotfile 文件中。用于绘制图形文件的算法根据操作系统和 DRAWGRAM 或 DRAWTREE 程序的选项来确定。如果有 PostScript 打印机并且选择 LaserWriter，你仅需把 plotfile 文件拷贝到打印机(UNIX：`lpr-Pprintername plotfile`)。

12.3.8 距离算法

距离算法计算改变的总数，按比对中的每对序列之间改变的类型记分。结果写入到一个距离矩阵用于构建树。距离算法计算代表序列之间改变量的分支长度。在距离计算中，间隙按未知字符记分而被有效地过滤掉。去除模棱两可的比对或未改变的字符会影响分支长度的估计。图 12.6 所示的流程图说明了操作步骤和所需要的文件。采用下列步骤构建一个距离树。

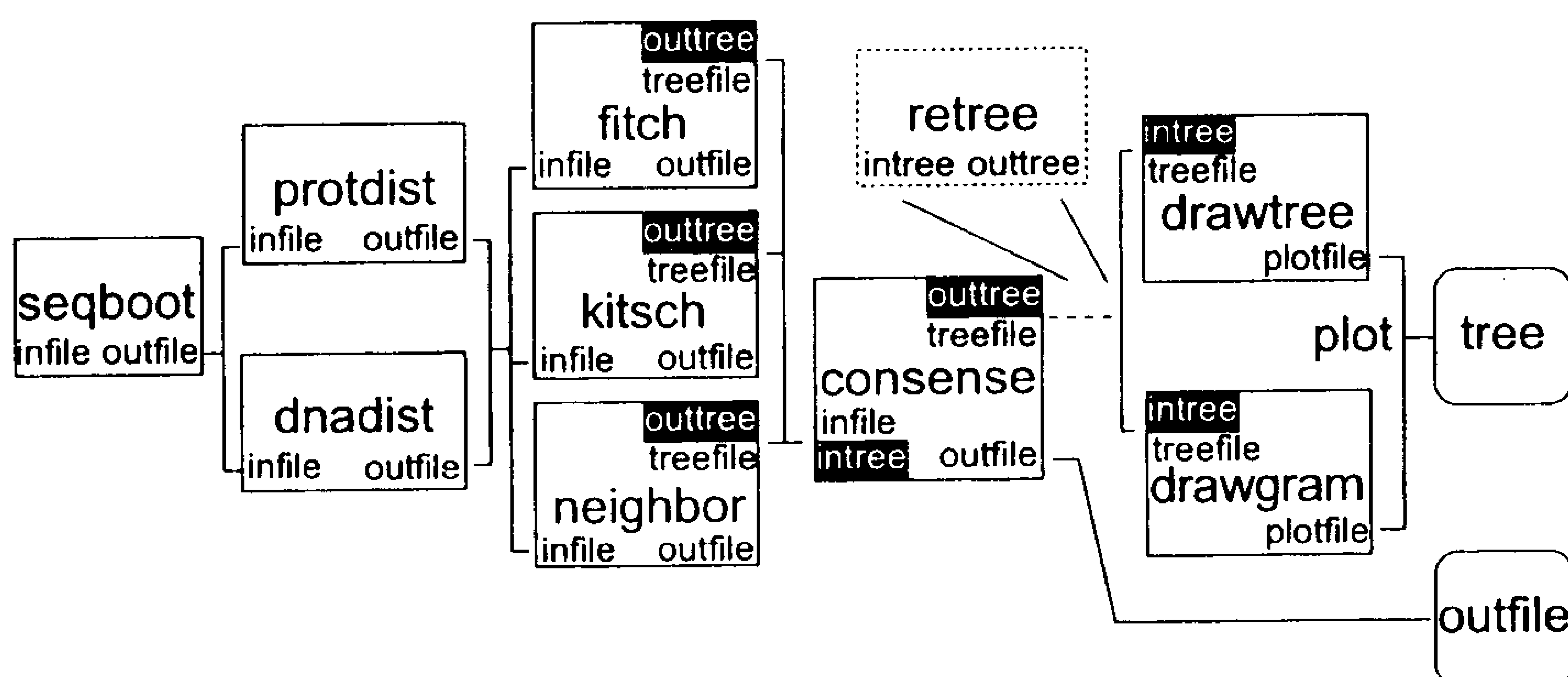


图 12.6 生成距离树的流程图

用法和图 12.4 相同，只多出一个步骤，即在用 FITCH、KITSCH 或 NEIGHBOR 程序构建树之前生成一个矩阵

(1) 拷贝 PHYLIP 格式的比对序列到 infile 文件中，此例中序列比对文件名称为 yourfile(UNIX：`cp yourfile infile`)。见 12.3.3 节的文件格式说明。

(2) 可选地，如果要采用解靴带算法数据包运行 SEQBOOT(UNIX：`seqboot`)。提供一个随机数目并接受缺省值生成 100 个数据包(见 12.3.4 节)。把 SEQBOOT 的 outfile 文件拷贝到 infile(UNIX：`cp outfile infile`)。注意如果采用解靴带算法，则会丢失分支长度。可以在单一的树上不用解靴带算法找出分支长度。大多数树把解靴带值按刻度距离树结合在一起就很方便。更复杂的树可以运行 FITCH 指定解靴带树作为用户树，具体如何指定用户树请参考程序说明。

(3) 为 DNA 数据包或蛋白质数据包创建距离模型可以运行 DNADIST 或 PROTDIST 程序(DNA 数据的 UNIX：`dnadist`，蛋白质数据的 UNIX：`protdist`)。有几个选项：

a. 选项 P，在 PROTDIST 程序中用户可以选择不同的变化模式。PAM，缺省值是很好的经验模式[9]。Kimura 算法的计算速度快，但有一些折中之处[10]。分类模型把氨基酸残基按功能分类。缺省的 Dayhoff PAM 矩阵虽然运行起来慢，但是一个更保守的选择。Kimura 双参数模型非常简单，在转换和颠换之间有 1~2 范围的比率，此模型通常很好用，但也有其他选择，如 Jin/Nei ML(最大似然法)和 J-C(Jukes 和 Cantor)。最大似然法模型和 DNAML 中使用的模型相似，速度也很慢。

b. 选项 T，设置转换/颠换比率，只在 DNADIST 中有，可以指定自己的比率。缺省值为 2，很好用。

c. 选项 C 和 W(PHYLIP 3.6)也只在 DNADIST 中有，用于指定可以按不同速率改变的区域。

d. 选项 M 指定用于解靴带算法的序列套数。

e. 选项 L，I，0，1 和 2 是常用的系统选项，不影响分析。缺省值在大多数情况下都适用。

(4) 拷贝 outfile 中的距离矩阵到 infile 中(UNIX：`cp outfile infile`)。

(5) 用 FITCH、KITSCH 或 NEIGHBOR 程序构建树。KITSCH 和 FITCH 相同，但采用了分子钟并且只计算有根树。NEIGHBOR 是构建树的基本程序，运行速度快。各程序的命令都相似。下列的例子是 FITCH[11]程序的：(UNIX：`fitch`)。

a. 选项 D，最小进化(PHYLIP 3.6)——此方法可用于当树中有负的分支长度时的手动方法。设置为 Fitch-Margoliash。

b. 选项 U 可以指定自己的树。文件格式参考 PHYLIP 说明文件，可以计算吝啬树然后采用距离树的分支长度。

c. 选项 P 设置公式的指数，计算标准误差时设置为 2。

d. 选项-，一些树可能生成负的分支长度，采用缺省值可以使所有的负值分支长度变为 0。

e. 选项 O 可以指定一个输出组，输出组的选择细节见 12.3.1 节第(2)项。

f. 选项 G 允许全部重排。每一次分组都去除然后重新加入到树中，以保证所有的安排都考虑。

g. 选项 J 使序列随机化。同 12.3.6 节(3)b 项。

h. 选项 M 指定用于解靴带算法的序列套数，如果要采用解靴带数据记得设置此选项。

i. 选项 I 指定缺省的 interleaved 序列格式，图 12.2 是序列格式。

j. 选项 L，R，0~4 是常用的系统选项，不影响分析。缺省值在大多数情况下都适用。

(6) 如果准备使用解靴带数据，则把树文件拷贝成 infile (UNIX：`cp treefile infile`)(PHYLIP 3.6：拷贝树文件成 intree，UNIX：`cp outtree intree`)。然后运行 CONSENSE(UNIX：`consense`)。CONSENSE 生成的 outfile

含有解靴带值。

(7) 用 DRAWGRAM 或 DRAWTREE 绘制树。这些程序需要由 FITCH、KITSCH 或 NEIGHBOR 程序以及采用解靴带数据时的 CONSENSE 程序生成的单个树文件。(PHYLIP 3.6：拷贝树文件成 intree，UNIX：`cp outtree intree`)见 12.3.6 节第(5)项的例子。DRAWGRAM(UINX：`drawgram`)。DRAWGRAM 或 DRAWTREE 的结果输出到 plotfile 文件。

(8) 用于绘制图形文件的算法根据操作系统和 DRAWGRAM 或 DRAWTREE 程序的选项来确定。如果有 PostScript 打印机并且选择 LaserWriter，你仅需把 plotfile 文件拷贝到打印机(UNIX：`lpr-Pprintername plotfile`)。

12.4 分析结果的翻译

(1) 树是从碱基开始的并沿着分支点或节点渐进的(图 12.7A、B 和 C)，程序 RETREE 是用来对树重新排序或改根。程序从 intree 文件中读入数据，并把结果写入 outtree 文件中。树文件(图 12.7E)可以容易地用文本编辑器编辑。

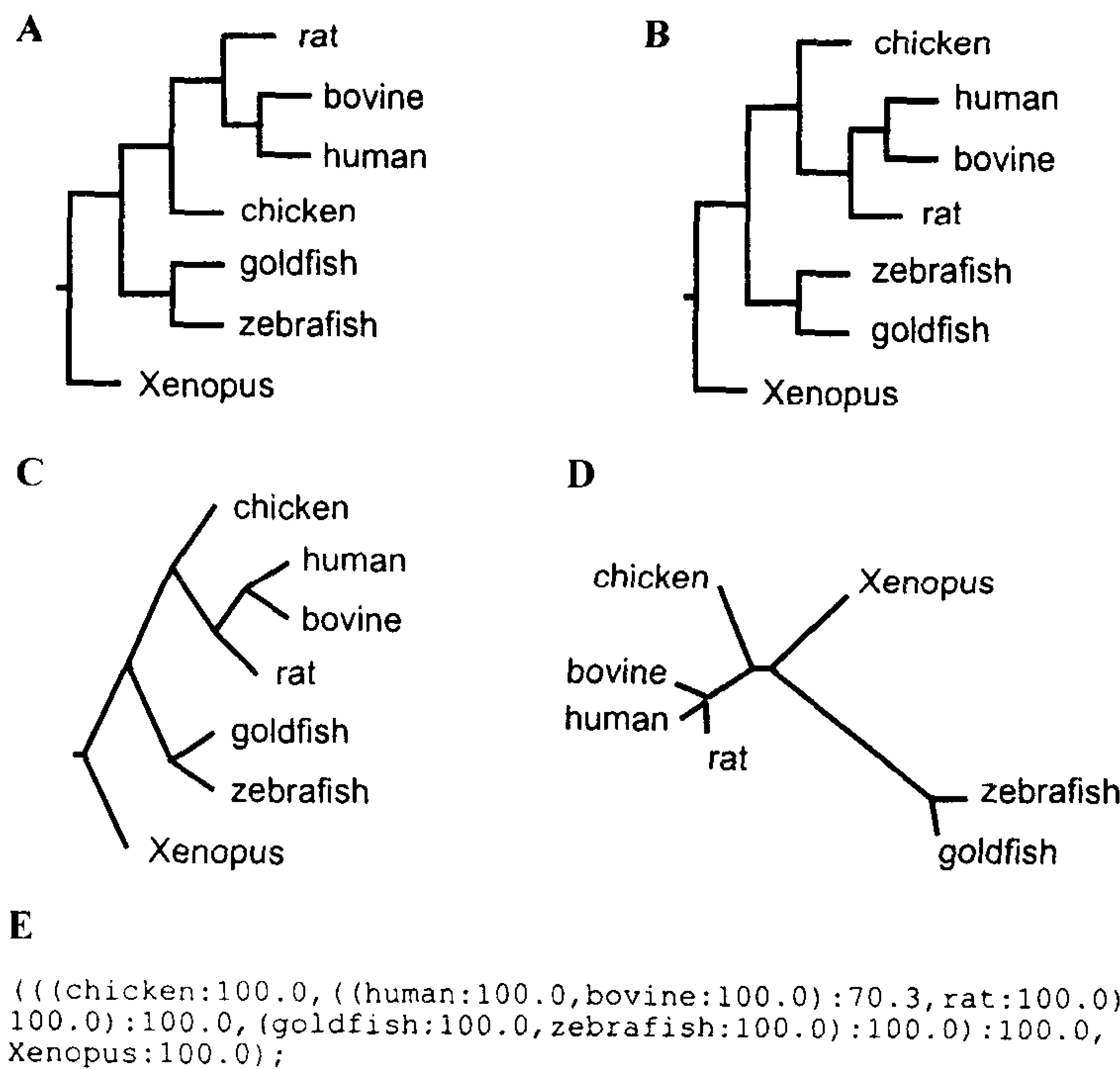


图 12.7 不同的树

A.和 B. phenograms。C. cladogram。A，B 和 C 是同一树的不同形式，由 DRAWGRAM 程序生成。D. DRAWTREE 程序生成的无根距离树，带分支长度比例。含有鱼的长分支代表序列中和其他动物之间的大量不同。E. 用于生成 C 的树文件

(2) 树的构建通常依赖于输入序列的次序。输入次序的效果可以运行程序的 J 选项来测试，把 J(jumble)设置为≥10，然后比较新树和原来的树。

(3) 采用适合于数据的绘图方式。尽管无根树可通过 DRAWGRAM 的 2p 选项绘制成 phenograms，但是用 DRAWTREE(图 12.7D)绘图则更易懂。

(4) 如果树生成的分支解靴带值范围是 90~100 之间，则结果是有统计学意义的，实际上分析树的每一种方法都会给出相似的结果，尽管实际的解靴带值会不同。当解靴带值低时，有助于试验几种不同的方案。

(5) 对于所有的树，建立一个距离树同时建立至少一种基于字符的树，例如，吝啬树或最大似然树这样比较保险。

(6) 通过许多不同统计学可信度方法生成相同的树很重要。例如，适于 51/100 个字符的树，每一种方法都可以生成。这样会产生高可信度的印象，但实际上真实的可信度约 50。

12.5 把结果传送到其他程序

因为结果是先写入到文件中，传送图文件(plotfile)到大量的图形程序是很简单的。选择 PostScript 选项把图文件输入到像 Adobe Photoshop®、Microsoft Word®、Macromedia FreeHand®程序中。编辑 PostScript 文件通常不可能。如果选择 HPGL 作为打印机格式，则图形文件以惠普的绘图仪描述语言生成。大多数基于 Windows 的程序，如 Macromedia FreeHand®程序可以把这些文件作为矢量文件输入，并且可以编辑，通常是做一些文字替换。在这些程序中线条宽度和颜色可随意改动。PHYLIP 3.6 还能生成在 Windows 中广泛使用的 bitmap 图形和 PICT 文件格式的图形。

(阮承迈　译)

参 考 文 献

[1] Li, W. -H. and Grauer, D. (1991) *Fundamentals of Molecular Evolution.* Sinauer Associates, Sunderland, MA.

[2] Ridley, M. (1993) *Evolution.* Blackwell Scientific Publications, MA.

[3] Li, W. -H. (1997) *Molecular Evolution.* Sinauer Associates, Sunderland, MA.

[4] Higgins, D., Thompson, J., Gibson, T., Thompson, J. D., Higgins, D. G., and Gibson, T. J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22,** 4673-4680.

[5] Felsenstein, J. (1985) Confidence limits on phylogenics: an approach using the bootstrap. *Evolution* **39,** 783-791.

[6] Fitch, W. M. (1971) Toward defining the course of evolution: minimum change for a specified tree topology. *Sys. Zool.* **20,** 406-416.

[7] Felsenstein, J. (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Molec. Evol.* **17,** 368-376.

[8] Kishino, H. and Hasegawa, M. (1989) Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from sequence data, and the branching order in Hominoidae. *J. Molec. Evol.* **29,** 170-179.

[9] Dayhoff, M. O. (1979) *Atlas of Protein Sequence and Structure,* volume 5, suppl. 3, National Biomedical Research Foundation, Washington, DC.

[10] Kimura, M. (1980) A simple model for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J. Molec. Evol.* **16,** 111-120.

[11] Fitch, M. and Margoliash, E. (1967) Construction of phylogenetic trees. *Science,* **155,** 279-284.

13 使用 Genotator 注释序列数据

Nomi L. Harris

13.1 引言：什么是 Genotator?

随着序列数据呈指数增长，越来越清楚地呈现出需要自动化的方法来帮助生物学工作者进行序列注释的工作。许多研究人员开发了分析 DNA 序列的工具，但是运行多个工具并解释其结果会很单调乏味，还易于混淆。

Genotator[1]是一个自动注释序列的工作平台，它提供了一种灵活明晰的系统来自动对基因序列运行一系列序列分析程序，同时还具备图形界面，能够使用户观察所有自动产生的注释和添加自己的注释。Genotator 的显示能够使用已注释的序列在多种细节水平上得到检查——从整条序列到单个碱基。通过显示各种类型序列分析的比对结果，Genotator 还提供了鉴别有意义区域(如可能的外显子)的直观方法。

Genotator 由两个主要部分组成，一个后端(back end)和一个浏览器。后端部分对 DNA 序列运行一系列的序列分析工具，处理所有的输入和输出格式。由 Genotator 运行的分析工具包括 5 个不同的基因查找程序、3 个同源性搜索程序和搜索启动子、剪接位点和可读框(ORF)的程序。Genotator 后端所得到的分析结果能用交互式图形浏览器观察，浏览器在画板上显示彩色编码的序列注释，可以滚动和缩放，能够在不同的细节水平观察已注释的序列。用户可以在一个独立的窗口中看到实际的 DNA 序列；当在图形显示中选择了一个区域后，序列显示就自动地加亮，反之亦然。用户能交互式地添加个人注释来标记目标区域。其他功能还包括引物设计和模式搜索，还能够检索读取 GenBank[2]中与注释序列有同源性序列的数据记录。

Genotator 在 UNIX 工作站上运行，后端用 perl 和 Tkperl 语言编写，调用各种序列分析程序。前端也由相同的语言编写，采用 Gregg Helt 的 bioTkperl 工具箱[3]。Genotator 是在 SUN 上开发的，也已经安装在 SGI 和 DEC Alpha 上。

Genotator 并非唯一可得到的序列注释工具软件(其他在参考文献[1]中讨论)，但是易于使用且可以免费获取，还给出了源程序(这样就可以改成更适合你的需要)，无需相关的数据库。

13.2 获得 Genotator

目前，全世界有几十个研究人员正在使用 Genotator。学术单位可以免费得到 Genotator，请与 Nomi Harris(`nlharris@lbl. gov`)联系获得 Genotator 的有关信息，`http：//www-hgc.lbl.gov/inf/genotator/need.html` 列出了 Genotator 所需的程序，并说明了如何获取这些程序。有些包括在 Genotator 软件中，其他的则必须和相关的作者联系。

13.2.1 安装 Genotator

注册获取 Genotator 后，就可以授权下载和安装。安装 Genotator 很简单：运行软件包中的 install-genotator 命令脚本即可，而安装 Genotator 所需的程序则较复杂。浏览器需要 perl5 和 Tkperl。后端能和它所知道的任意序列分析程序一起发挥作用，必须获取和安装你要后端使用的任何程序和序列数据库。

本章剩余的部分将讨论在安装好了 Genotator 之后如何使用它。

13.3 运行 Genotator 后端

Genotator 后端对序列文件进行一系列分析，并把结果保存，供以后浏览。在许多可以得到的序列分析工具中，选取了一部分整合到 Genotator 中。由 Genotator 调用的分析程序分为 3 个主要类型：基因查找程序(Genie[4]、GRAIL[5]、GeneFinder[6]、xpound[7]和 GENSCAN[8])、数据库同源性搜索(BLASTN[9]与人类或果蝇数据库重复序列数据库比较，然后其用 xblast[10]作为界面；基于 dbEST 的 BLASTN[11]；基于 GenPept[2]的 BLASTX[9])和序列特征预测工具(起始/终止密码子、可读框、启动子[12]及剪接位点[13])。Genotator 为每一分析程序提供了合适的输入格式，并且把输出翻译成一种可读的文本格式，就像 ACeDB[14]所采用的那种。输出文件按序列和用户层次进行组织。

Genotator 后端可以通过图形用户界面(GUI)或命令行选项激活，GUI 方式见图 13.1，要通过 GUI 激活 Genotator(假定已经在/home/genotator 处安装了 Genotator)，敲入：

```
/home/genotator/genotator
```

如果知道要注释的序列文件名称，可以在命令行上同时输入，如：

```
/home/genotator/genotator humtfpb
```

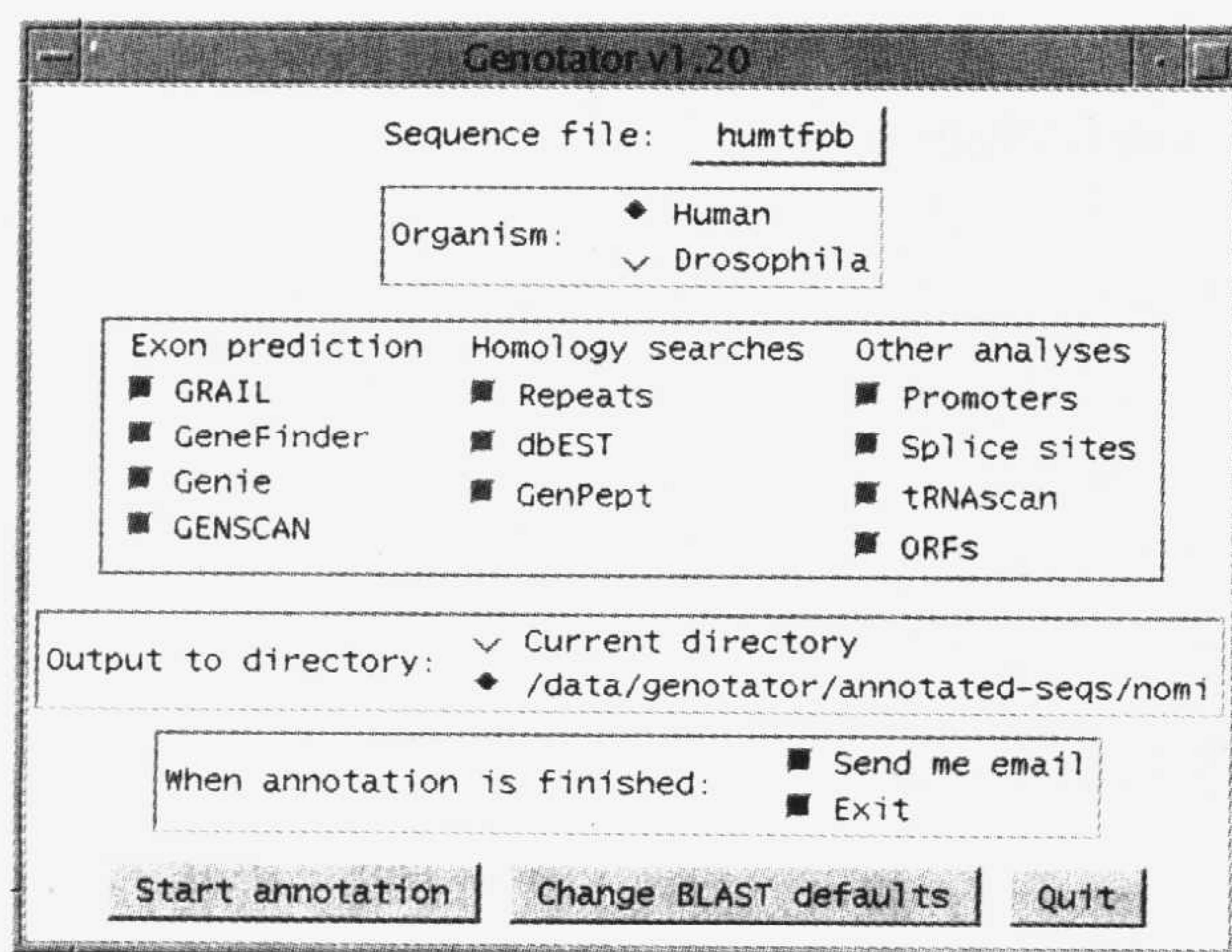


图 13.1　Genotator 后端的图形用户界面

用户能选择要注释和分析的序列

13.3.1　序列文件格式

选择要注释的序列文件时，点击文件选择框(这里标记为“humtfpb”，因为用户已经选择了 humtfpb 作为欲注释的序列)，出现一个文件选择菜单。序列文件可以接受的格式为：

- 纯文本(只有序列，没有行号或其他东西)。
- FASTA(一个以“>”起始的标题行，后跟序列行)。
- 类似 FASTA 格式，在标题行以“；”起始，而不用“>”起始。
- GenBank 格式。

以下是一个 FASTA 格式的例子：

```
>gb|J02846|HUMTFPB Human tissue factor gene,
complete cds.
GAATTCTCCCAGAGGCAAACTGCCAGATGTGAGGCTGCTCTTCCTCAGTCACTAT
CTCTGGTCGTACCGGGCGATGCCTGAGCCAACTGACCCTCAGACCTGTGAGCCGA
GCCGGTCACA[etc.]
```

13.3.2　Genotator 选项

可以通过 Genotator 配置的选项包括：

- 生物(人类或果蝇；默认为人类)。
- 进行何种分析(默认为全部；去掉打勾的是不准备进行的分析)。
- 结果保存目录[默认为所有注释序列存放的目录(按用户名分子目录)，但是用户可能会希望建立自己的 Genotator 目录，并将注释存储在自己的当

前目录中]。

- 当分析过程完成后，是否希望通过 e-mail 通知(默认为是)。
- 默认的 blast 取舍点(cutoff)值(点击 Change BLAST defaults 键改变此选项)。在进行其他 BLAST 搜索前，还能控制是否将重复序列(如 Alu)过滤掉。

13.3.3 批处理模式运行 Genotator

如果想对一组序列文件运行 Genotator，用命令行选项的批处理模式会更容易。通常，只需键入如下一类的内容：

```
/home/genotator/genotator-batch seq1 seq2 seq3...
```

其中 seq1 等是文本格式或 FASTA 格式的序列文件，`-batch` 告诉 Genotator 不要打开 GUI。

以-h(help)选项运行 Genotator 时，会打印出一长串的命令行选项：

```
Usage: genotator[seqfile1[seqfile2...]]
[-human or-drosophila][-none][-nomail][-exit]
    [-d(ebug)][-noblast][-nomask][-dir
    output_dir][-exon][-all][-batch][-grail]
    [-genefinder][-genie][-genscan]
    [-xpound][-genemark][-genpept][-est]
    [-repeats][-promoters][-splice][-trnascan]
    [-orf]
```

`[-h(elp)]`：打印此帮助信息。

`[seqfile1...]`：序列文件名(纯文本、FASTA 或 GenBank 格式)，可以指定多个序列文件。

`[-human or -drosophila]`：序列来源于哪种生物(默认为人类)。

`[-none]`：不分析选定框。

`[-nomail]`：分析完不发 e-mail(默认为发 e-mail)。

`[-exit]`：完成后退出(默认为不退出)。

`[-d(ebug)]`：调试模式(开发人员用)——打印 Genotator 要进行的工作，但并不真去做。

`[-noblast]`：试图恢复以前的 BLAST 输出结果，但是重做 BLAST 后处理。

`[-nomask]`：在 BLAST dbEST 和 GenPept 前去掉重复序列(默认为去掉)。

`[-dir output_dir]`：在 output_dir(的子目录)储存结果。

`[-exon]`：只运行基因查找程序(批处理模式下)。

`[-all]`：在批处理模式下运行所有分析。

`[-batch]`：在批处理模式下运行部分分析；由其他自变量指定具体分析项目。其他自变量是可以与`-batch` 选项一起指定的序列分析的名称。

13.4　Genotator 浏览器

13.4.1　激活浏览器

一个序列由 Genotator 处理过后，Genotator 浏览器提供了一个交互式的图形界面来观察注释结果。Genotator 浏览器能通过以被注释的序列文件名作为自变量来激活，如：

```
/home/genotator/genotator-browser humtfpb
```

如果没有加自变量，则会显示已经注释序列的列表，用户注释的序列放在前面。其他序列目录用“…”简略，如图 13.2 所示。要看简略目录中的序列时，双击目录名(如 liepe…)，然后是目录中的序列(或子目录)在前面显示。如果在列表中有许多序列名称，可以采用 Find 按钮帮助找到想要的序列。找到后，双击之(或单击并点击 Select)。选择列表会消失，浏览器载入相应的注释。

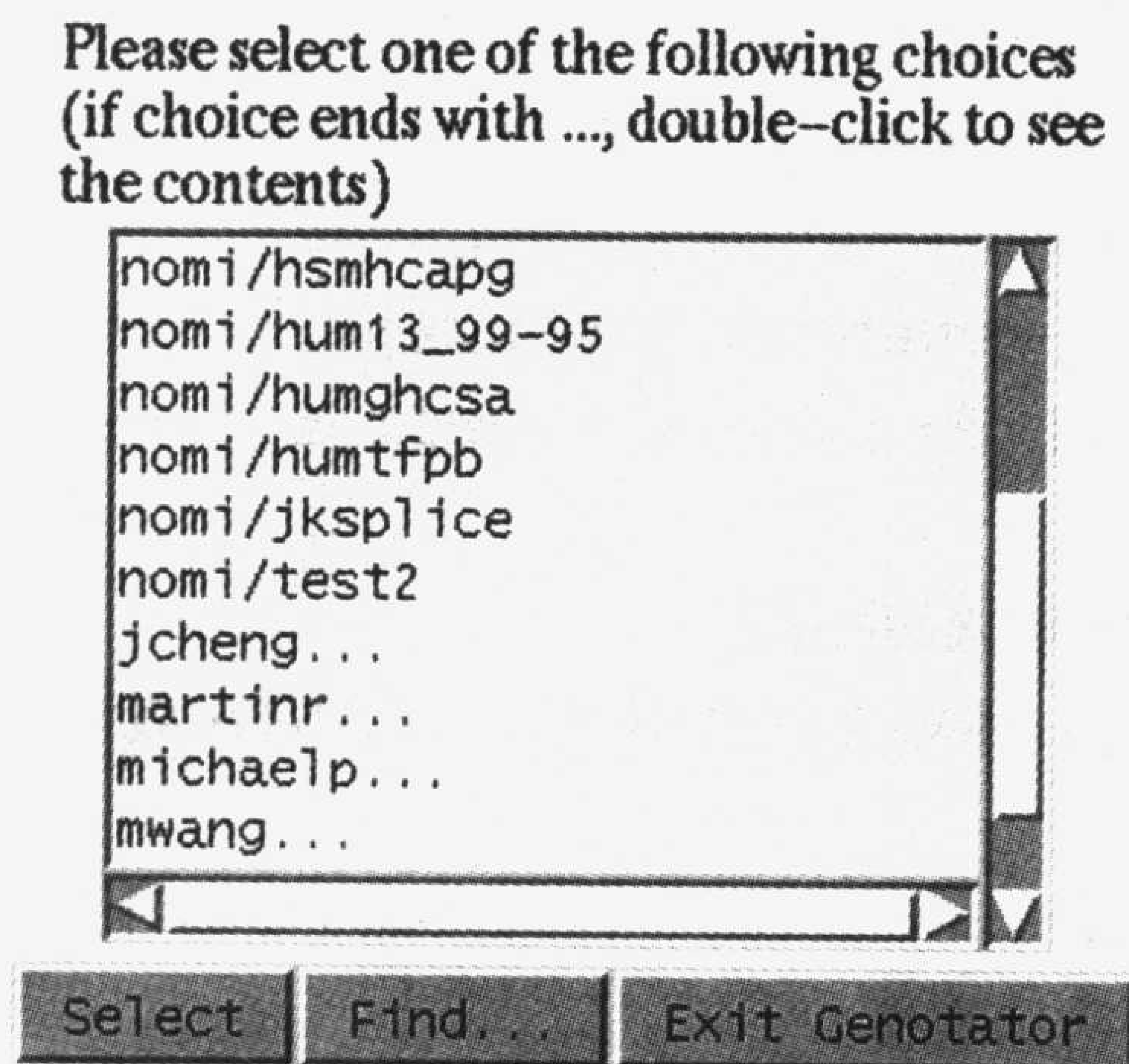


图 13.2　如果 Genotator 浏览器以无自变量方式激活，则出现已注释序列的列表。非当前用户的目录用“…”表示

13.4.2　图形显示

Genotator 的主要显示方式为图形显示(map display)。在图形显示的中心是代表序列的水平轴线，在轴线的上方是正向链的注释，下方是反向链的注释。沿轴线的数字代表 kb 碱基数。每一类注释(如 GRAIL 外显子)显示在自己的行上，用其本身的颜色。显示的结果可以缩放和滚动以便仔细观察感兴趣的区域。缩放时用鼠标拖拽 ZOOM 棒，或把光标放在其旁边然后单击来逐渐缩放。滚动时拖拽位于图形显示下方的滚动棒(图 13.3)。

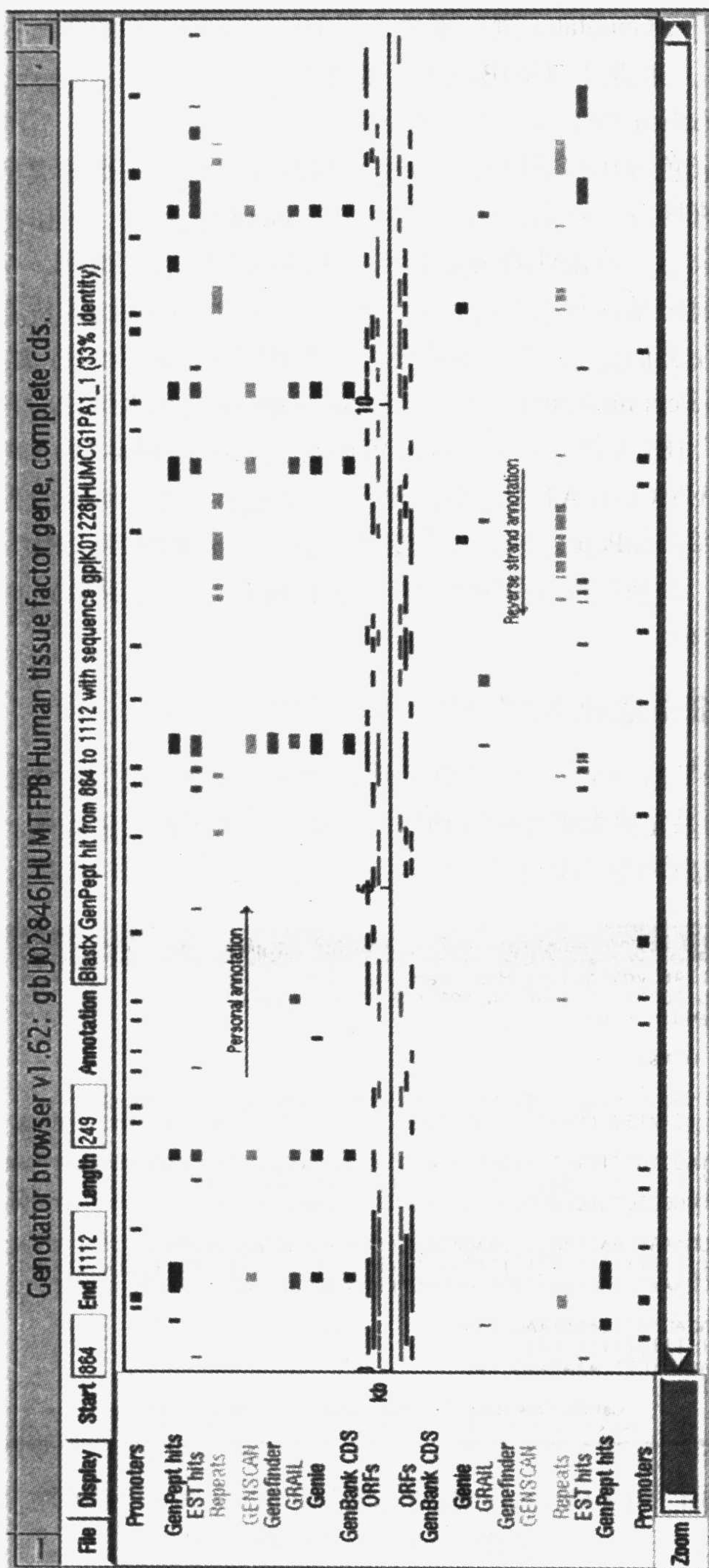


图 13.3　Genotator 浏览器提供了序列注释概览，正向链和反向链上的注释用中轴上下的色块表示

在图 13.3 中，Genotator 浏览器显示了 HUMTFPB[15]的注释，这是一个人组织因子基因序列，来源于 GenBank(彩图可以在 http://www-hgc.lbl.gov/homes/nomi/chapter.html 上看到)。

图中每一彩色方块代表已经注释的序列区域，每一注释的类型由方块的颜色加上不同的行来区分。点击左边的行标签(如 GenPept hits)，可以得到这行中的注释类型的更多信息。点击注释方块时在该方块的四周加上黑框，并在浏览器顶部的文本窗口中显示某注释附加的信息，包括注释的起始和结束位置，可能会有一个分值和其他相关信息。例如，如果点击一个 BLAST 命中(hit)，文本窗口就显示："BLASTX GenPept hit from 864 to 1112 with sequence gp | K01228 | HUMCG1PA1_1 (33% identity)" [BLASTX 在 GenPept 库中找到和序列 864~1112 部分相同的序列 gp | K01228 | HUMCG1PA1_1(33%同源性)]。简洁地描述了所用的程序(BLASTX)、所查找的数据库(GenPept)、所命中的数据库序列(gp | K01228 | HUMCG1PA1_1 是其 GenPept ID)、与数据库序列相似的区域(第 864~1112 个碱基)，以及命中的一致性的百分比(33%)。

13.4.2.1 更细微地观察 BLAST 命中结果

双击 BLAST 命中结果可以更详细地对其进行观察，(注意：Tkperl 需要快速而仔细地双击，如果先放大把光标放到注释方块的中央可能会更好些。)对于 BLASTN 命中结果(核苷酸序列)，能在另一个窗口显示完全的比对(图 13.4)，能存储或打印。

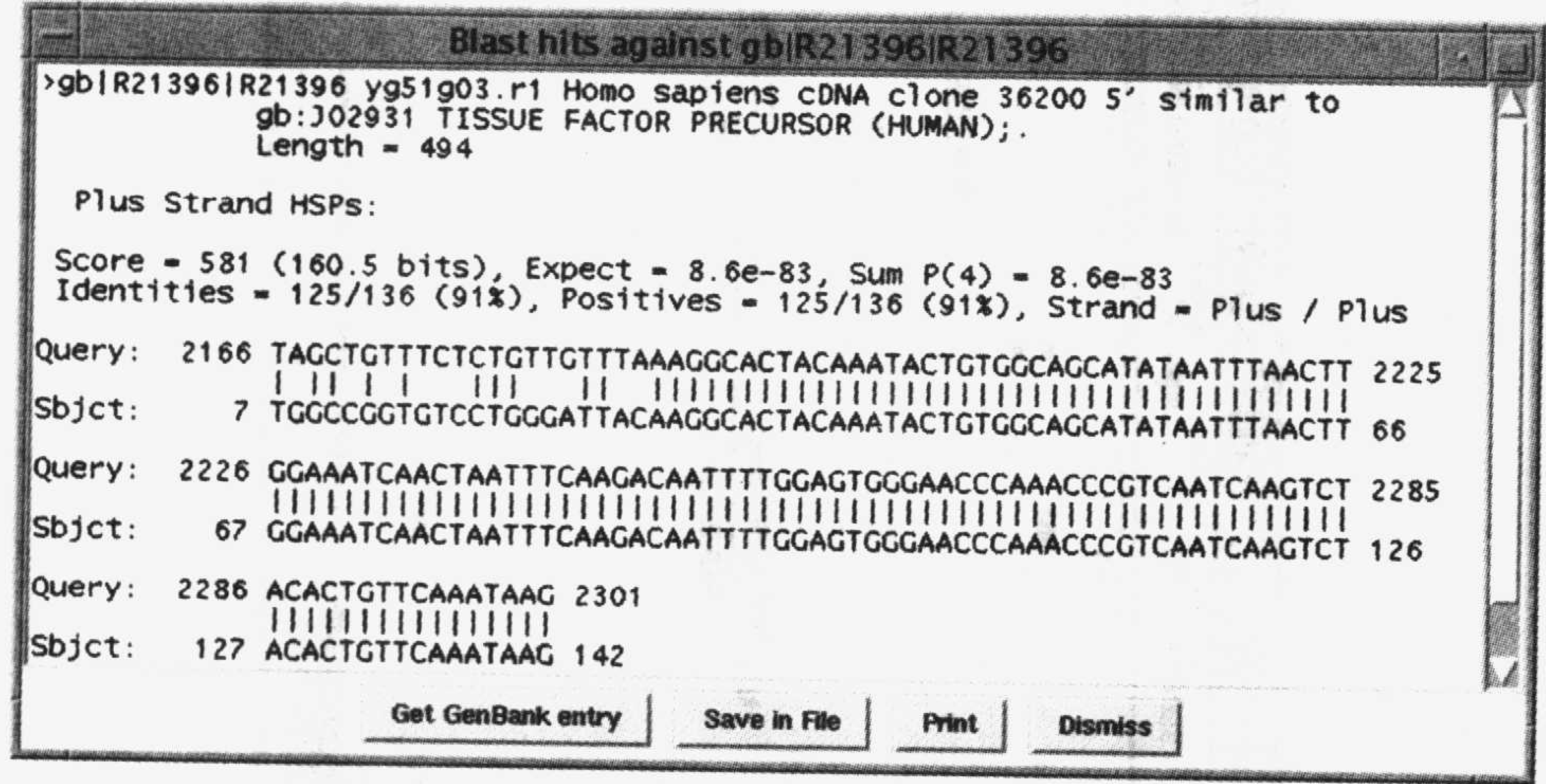


图 13.4 当在 Genotator 窗口中双击一个 BLASTN 命中结果时，弹出这样一个窗口，显示实际的比

当在 Genotator 显示上双击 BLASTX 在 GenPept 的搜索命中结果时，会弹出一个 Sanger 中心的 BLAST 命中结果的窗口 Blixem[16](图 13.5)。

图 13.5　当双击 Genotator 显示的 BLASTX 搜索 GenPept 结果后，弹出专门显示 BLAST 结果的 Blixem 程序界面

Blixem 用水平粗线代表靠近点击区域的命中结果，线的垂直位置表示一致性的百分比。蓝框显示在下方扩展的区域，代表实际的比对结果，可以用鼠标中间键移动蓝框，因为 BLASTX 把 DNA 序列与一个氨基酸数据库进行比较，命中结果用所有的3种读框来显示。精确和类似的匹配用颜色加亮显示。要关闭 Blixem 窗口，在窗口的空白灰色区域点击鼠标右键，会弹出一个菜单，其中一个选项是 Quit。

13.4.2.2 Genotator 浏览器功能：文件菜单

Open：选择欲显示的注释序列。

Reload：重载当前序列(通常在错误地添加或删除了个人注释时有用)。

Print：把图形显示作为 PostScript 文件截取并(可选)送到默认的打印机上。(注意：此项功能并非在所有的系统上都有——打印需要 xwd 和 xwd2ps。)改变默认的打印机时，在打印前设置 PRINTER 环境变量：

```
setenv PRINTER myfavoriteprinter.
```

Summary report：产生描述全部注释的长文本报告，可以存储和/或发送到默认的打印机上。

Submit comment：可以把评论直接提交到 Genotator 开发者手中。

Output selected region：选中任意区域存储到磁盘(FASTA 格式)以便进一步分析。

Quit：退出浏览器。如果添加或删除了自己的注释，会询问是否要保存改变。

13.4.2.3 Genotator 浏览器功能：显示菜单

Hide/Show complement：在图形显示中隐藏(或显示)互补链(在中心轴的下方)。如果你只想保存正向链的图形时很有用。

Display sequence：在另一个窗口中显示正链碱基，在下一节中有详细讨论。

Display sequence complement：显示互补链碱基。

Show/hide splice sites：当浏览器打开时，不显示剪接位点，使用本功能可以将其显示。

Show/hide start/stop codons：当浏览器打开时，起始/终止密码子不显示，使用本功能可以将其显示。

Delete selection：在选择框内从屏幕上删除所有的注释。(注意：这些注释并非永久性地删除——只有个人注释才可以永久删除。)如果重载序列，所有自动产生的注释会重新出现。删除功能在做幻灯片时最有用。

Get GenBank record for hit：此命令仅当点击了一个 EST 或 GenPept 命中结果后才能使用，它查询 GenBank 数据库，试图找到与目标序列(命中的序列)匹配的记录。如果它能找到 Netscape，则在 Netscape 中会显示 GenBank 记录，否则，打

开一个新的文本窗口。(注意 Netscape 启动需要些时间。)有时不能显示 GenBank 记录，因为 Genotator 找不到目标序列的适当记录。

Personal annotations...：出现一个控制窗口，以添加和删除个人注释，这在 13.4.4 节中有进一步叙述。

Design primers：对选定的区域进行引物设计(见 13.4.5 节)。

13.4.3 序列显示

图形显示表明了整个序列的概览。Genotator 浏览器还能在一个独立的窗口中显示实际的 DNA 序列(或其互补链)，见图 13.6。当用户在一个图形显示窗口中选择一个注释时，相应的区域就会在序列显示窗口中以适当的颜色加亮。例如，在图 13.6 中，在图形显示窗口中所选中的 GenPept 命中结果在序列显示中被加亮。

图形显示的序列显示之间的相互作用是双向的：当在序列显示窗口中选中一个区域时，则在图形显示窗口中加上选框。(如果光标在两行之间的中途时碰巧松开鼠标，当在序列显示窗口划出一块区域时，在图形显示窗口的加框区域会错误地从零开始，这是一个已知的 Tkperl 隐错，在 Genotator 内无法纠正。)如果正链显示在序列窗口中，那么点击正链中的注释时会在序列窗口中将其加亮，但点击反向链的注释时就不会。

13.4.3.1 序列显示功能

Show complement/forward strand：在显示正链和互补链之间切换(这会使程序先关闭显示窗口后又重新打开一个)。请注意为了和图形显示兼容，当显示互补链时，不是反向互补。当显示正链时，为了从视觉上提醒当前显示的是哪条链，序列显示窗口中的棒颜色为浅蓝，如果显示互补链，棒为粉红色。

Sequence highlights：此选项控制是否加亮新序列(如当在图形显示窗口点击一条注释时出现的那些新序列)替换原有的加亮部分，或将其合在一起(默认值是“替换”)。这在黄色的加亮上，即当用鼠标划出一段序列显示区域时无效，因为这样总会替换原有的加亮。如果要清除所有的加亮区域，选择“替换”复选按钮，然后在序列显示窗口中点击鼠标(不要拖拽)。

1) 文件菜单

Output selected region：在图形显示窗口使用此功能时，能把选择区域输出到一个文件。如果在互补链上选择了一个区域，则会存成反向互补形式(即使序列窗口中显示的序列不是反向的)。

Print this window：把序列窗口保存为 PostScript 和(可选)送到默认的打印机上(如果有 xwd 和 xwd2ps)。

Close：关闭序列窗口。

图 13.6　序列显示

用户在图形显示窗口中点击注释时，在序列窗口中的相应部分就会加亮

图 13.7　模式搜索例子

2) 显示菜单

Find pattern...：此选项弹出一个窗口，让用户键入一个字符串或常规表达式，以在序列中搜索哪一个。字符串是 A、C、T、G 组成的任意序列，如：CCGCGTTG，也可以表示限制酶位点或基序等。还能搜索 UNIX 格式的常规表达式，例如，假设要查找所有 A 后为 C 或 G，然后是一个或几个 T，再后面为 A 的序列，对这种情况的 UNIX 格式的常规表达式为 A[CG]T+A。在图 13.7 中，Genotator 找到了并且加亮了所有与之匹配的子序列。有关 UNIX 格式的常规表达式的信息可以通过键入 man reguexp 命令或在地址：http://www.wiley.com/compbooks/unixshell/appendix-i.html 得到。

Personal annotations...：打开添加或删除个人注释的控制面板，如下一节所述。

Highlight stop codons：在一个或所有 3 个读框中加亮终止子，以彩色加框表示。

13.4.4 添加个人注释

Genotator 浏览器允许用户在图形或序列显示中添加新注释，这些个人注释和以前计算机加的注释一起保存。图 13.8 表明了添加或删除个人注释的界面，要往图形或序列显示中添加个人注释，用户选择序列中的某一区域，在文本框中键入注释的文本，然后点击 Add Annotation to Map 或 Add Annotation to Sequence。当创建了个人注释后，可以指定其颜色，点击 forestgreen 按钮弹出一个色彩选择菜单。改变注释颜色只会影响将要添加的注释，以前添加的不会改变，这样可以用不同的颜色表示不同类型的注释。

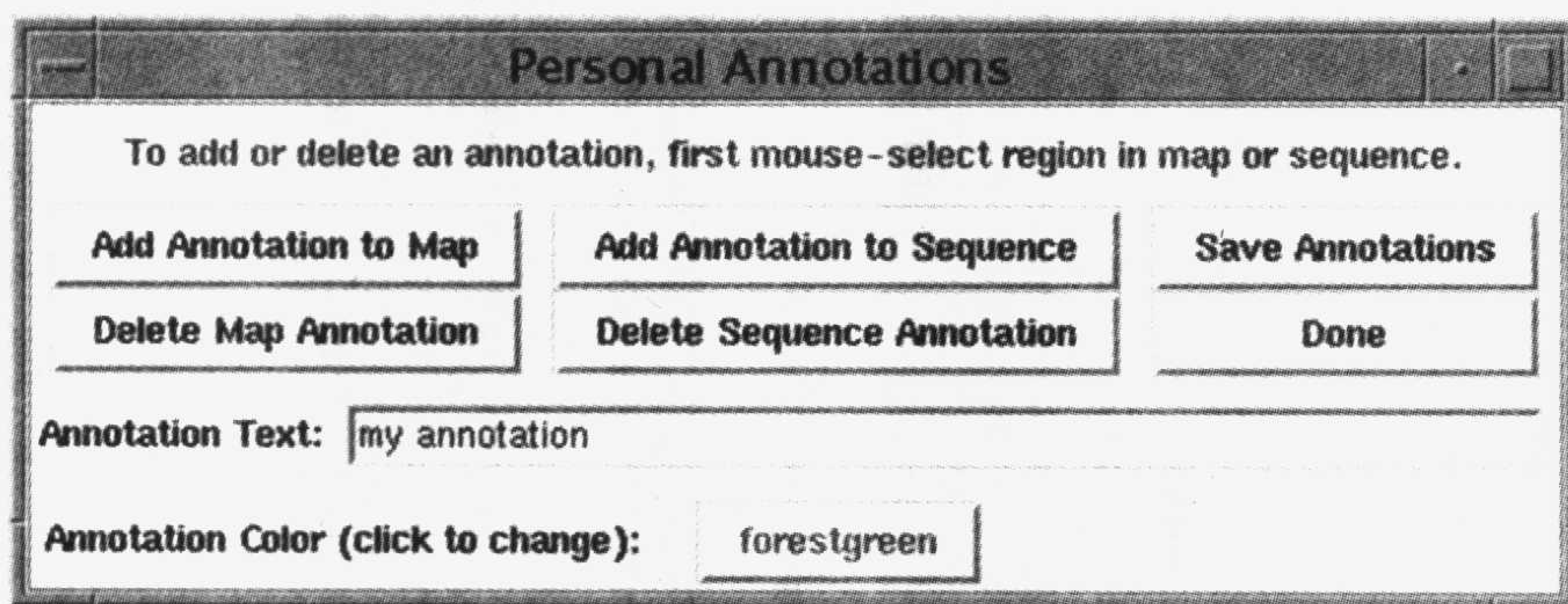


图 13.8 添加注释界面

涉及大部分序列的注释通常添加到图形窗口中，而涉及小区域(如一条引物)的注释则添加到序列中更合适。所有个人注释将和自动产生的注释一起存入数据库。个人注释的例子见图 13.3(“个人注释”和“反向链注释”)的图形显示和图 13.6(“序列中的个人注释”)的序列显示中。

13.4.4.1 删除个人注释

要删除图形显示中的个人注释，用鼠标在你要删除的注释周围选择一个框(如果先放大则会容易操作)。注释会在显示窗口中消失，但是，在点击 Save Annotations 之前，它并不是永久消失(如果不存盘退出浏览器，会询问你是否保存个人注释)。这里没有删除的 undo(恢复)功能，但是如果你误删除了一条注释，可以用 reload 把以前保存在数据库中的所有个人注释都恢复(当然，尚未保存的任何新注释将丢失)。

由于序列显示窗口中的个人注释可能会重叠，将其删除的步骤略有不同。选择 Delete Sequence Annotation 会弹出一个窗口，显示序列注释的位置和标记。如果单击其中的一条，该注释会在序列显示中被加亮成青色。双击，或按下 Delete 键，会删除该注释。序列显示窗口会消失，并重新显示出不包括已删除注释的窗口。

13.4.5 引物设计

为帮助用户对某目标区域设计引物，Genotator 能调用 Primer3[17]程序。首先选择一段序列区域(用鼠标拖拽在图形或序列显示窗口中划出一段区域，或在图形显示窗口中点击一条注释)，然后在菜单中选择 Design Primers，按需要改变 Primer3 的任意默认选项，最佳的上下游引物就会在终端上打印出来(这样就可以切下并粘贴到引物订购单上)并且在序列显示窗口显示。有关 Primer3 的更多信息见第 20 章。

13.5 个性化 Genotator

前边的几节主要描述了帮助 Genotator 用户进行配置的选项，而 Genotator 还有程序员级的设置项目，有能力的 Perl 程序员应当能够改变或增加 Genotator 的功能。最容易的就是改变基因查找程序的选项——运行哪个程序，以何顺序显示。添加新的基因查找程序到 Genotator 软件包中涉及拷贝和修改一个已知的基因查找程序(如 GRAIL)的功能，以及编写一个编译器来把新的基因查找程序的输出格式转成 Genotator 能读的格式。这样将会使 Genotator 在另一个数据库上运行 BLAST 也同样快捷，例如，全部 GenBank 数据(这需要给结果选择一种颜色和平衡位置)。

致谢

感谢 Gregg Helt 的帮助，他编写了 bioTkperl 工具箱和图形浏览器(AnnotP1)，实现了 Genotator 浏览器中的许多功能；感谢 Martin Reese，编写了由 Genotator 调用的数个序列分析程序，帮助我调试了 Genotator 的早期版本，并对 Genotator

的早期论文提供了有深度的建议；感谢 Colin Collins，热情地支持在他的小组中使用 Genotator；感谢 Suzanna Lewis，Berkeley 果蝇基因组计划信息学小组的主任，我也在此工作；感谢 Judith R. Harris 建议使用 Genotator 的名称。

(李　鹏　李慎涛　译)

参 考 文 献

[1] Harris, N. L. (1997) Genotator: a workbench for sequence annotation. *Genome Res.* **7**, 754-762. To obtain *Genotator*, email the author, `nlharris@lbl. gov`.

[2] Benson, D. A., Boguski, M. S., Lipman, D. J., Ostell, J., and Ouellette, B. F. (1998) GenBank. *Nucleic Acids Res.* **26,** 1-7.

[3] Helt, G. (1997) Data visualization and gene discovery in *Drosophila melanogaster*, PhD thesis, University of California at Berkeley.

[4] Kulp, D., Haussler, D., Reese, M. G., and Eeckman, F. H. (1996) A generalized hidden Markov model for the recognition of human genes in DNA, in *Proceedings of the Conference on Intelligent Systems in Molecular Biology' 96,* AAAI/ MIT Press, St. Louis, MO., pp. 134-142.

[5] Xu, Y., Mural, R. J., Shah, M. B. and Uberbacher, E. C. (1994) Recognizing Exons in Genomic Sequence Using GRAIL II, in *Genetic Engineering: Principles and Methods,* vol. 15(Setlow, J., ed.), Plenum, New York, NY, pp. 241-253.

[6] Green, P. (1994) Ancient conserved regions in gene sequences. *Curr. Opin. Struct. Biol.* **4**, 404-412.

[7] Thomas, A. and Skolnick, M. H. (1994) A probabilistic model for detecting coding regions in DNA sequences. *IMA J. Math. Appl. Med. Biol.* **11**, 149-160.

[8] Burge, C. and Karlin, S. (1997) Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* **268,** 78- 94.

[9] Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990) Basic local alignment search tool. *J. Mol. Biol.* **215,** 403-410.

[10] Claverie, J. M. and States, D. J. (1993) Information enhancement methods for large scale sequence analysis. *Comp. Chem.* **17**:191-201.

[11] Boguski, M. S. (1995) The turning point in genome research. *Trends Biochem. Sci.* **20,** 295-296.

[12] Reese, M. G. and Eeckman, F. H. (1994) New neural network algorithms for improved eukaryotic promoter site recognition, in *The Seventh International Genome Sequencing and Analysis Conference,* Hilton Head Island, South Carolina, September 16-20, 1995.

[13] Reese, M. G., Eeckman, F. H., Kulp, D., and Haussler, D. (1997) Improved splice site detection in Genie, in *First Annual International Conference on Computational Molecular Biology (RECOMB)*, 1997, Santa Fe, Waterman, M., ed., ACM Press, New York, NY.

[14] Durbin, R. and Thierry-Mieg, J.(1991)A *C. elegans* Database. Documentation, code and data available from anonymous FTP servers at `lirmm.lirmm.fr,cele.mrc-lmb.cam.ac.uk and ncbi.nlm.nih.gov.`

[15] Mackman, N., Morrissey, J. H., Fowler, B., and Edgington, T. S. (1989) Complete sequence of the human tissue factor gene, a highly regulated cellular receptor that initiates the coagulation protease cascade. *Biochemistry* **28,** 1755-1762.

[16] Sonnhammer, E. L. L. and Durbin, R. (1994) A workbench for large scale sequence homology analysis. *Comput. Applic. Biosci.* **10**, 301-307.

[17] Rozen, S. and Skaletsky, H. J. (1996) Primer3. Code available at `http://www-genome.wi.mit.edu/genome software/other/primer3.html.`

14 低价位的凝胶分析系统

Jeffry A. Reidler

14.1 引言

能进行凝胶数据处理的设备有多种档次，从 300 美元的胶片相机到 15 000 美元的全套凝胶分析工作站不等。本文重点介绍低端凝胶数据处理设备的组成部分：图像采集、存储和分析模块。而工作站由以下模块组成：透射仪、照相机、采集卡、分析软件、存储设备和打印机。早期的处理系统仅由一个宝利来相机组成，不能转换成计算机能处理的数字信号。要把图像转换成数字信号还必须增加具有数字信号输出功能的照相机或摄像头，对某些应用还要加滤光片。现在以低于 2500 美元的价格就能组装一个工作站，包括数码相机、镜头、滤光片和分析软件，当然必须在实验室有透射装置和计算机的条件下。打印机可以用照片打印机，约 1300 美元，也可以采用廉价的彩色喷墨打印机。

14.2 设备

很多公司都经营相关的设备，如 BioRad、UVP、Ultra-Lum、Fotodyne、Stratagene、Alpha Innotech 等。标准的凝胶成像分析系统由以下几部分组成：透射仪、暗箱、数码相机、计算机、软件和打印机，价格从 6000~15 000 美元不等。在互联网上能找到这些资料和经销商，除能组装这些系统外还提供常规服务。

透射仪可以从仪器商店买到，价格在 1000 美元左右。在大多数情况下，实验室用的透射仪就能满足要求，一般透射仪只有紫外灯，可以通过放一块白光转换板把紫外光转换成白光。以下以 UVP 公司的透射仪来详细说明。

支架用于在暗室固定相机，由三脚架或钢柱组成。有标准紧固螺钉，大小为 1/4×20 英寸，适用于 35mm 的 SLR 相机。暗箱功能相似，没有暗室也能使用。

图像采集可以用实验室现有的黑白摄像头，加上图像采集卡组成，采集卡作用是把图像信息转换成数字信号，14.6 节有供应商的信息，图像采集软件用 Scion LG-3，能和苹果机和 PC 机的 NIH Image 或 Scion Image 兼容。

低照度相机也可以用于荧光凝胶，如溴化乙锭(EtBr)荧光图像的采集。数码相机具有高性能和高清晰度，如越来越便宜的 Kodak 相机 DC120/DC210/

DC260 等。采用低照度摄像头优点是速度快、价格低，典型的有 Scion 公司的 GMS300，1695 美元，白光和紫外光都适用。

解析度也是要考虑的因素。欧洲 B/W CCIR(PAL 制式)的视频解析度是 768×576 像素，比美国的 B/W RS170(NTSC 制式)的解析度 640×480 高出约 30%。一个 CCIR 制式的相机与插值技术结合，能使长焦距图像达到 1000 像素的解析度，能满足大多数凝胶分析的需要。如果要求更高，可以采用 DC260(2000×1600)，或更高清晰度的数码相机。

滤光片的使用视用途而定，从 Chroma Technology 和 Omega Optical 公司可买到适用于 EtBr 凝胶的三孔干涉滤光片。大多数还附带转接环，方便滤光片直接安装到变焦镜头上。这些干涉滤光片能防止紫外和红外光透过，但有 80%的所需波长光能穿透。把偏振相机换成凝胶成像系统时，可能需要换滤光片，因为偏振相机底片和摄像机的硅基质 CCD 的灵敏度不同。

镜头的焦距范围是 8~48mm，最长焦距时的最小视野为 2 英寸(35mm)，最大焦距时的最小工作距离为 14 英寸(35cm)。还可以选择焦距为 11~69mm 的镜头，在凝胶成像系统中，这类镜头常在暗箱中使用。Rainbow 和佳能都提供这类镜头，带 46mm 的安装螺纹；这类公司还有 Toyo 和 Cosmicar，价格在 250~500 美元之间。

较经济的选择是 12mm、16mm、25mm C-mount 的镜头，安装到支架上，可以根据不同的凝胶尺寸调节高度。在镜头和相机之间放置一个薄环，这些镜头即变成放大镜头。1mm 厚的环能使原来的视野缩小一半。这些镜头的价格从 80 美元到 150 美元不等。

如果是亮度低的凝胶，应该选择光通量大的镜头，或者在一定的焦距范围内 F 值最小的镜头。我们发现 25mm 的 F1.4 镜头，在 25mm 焦距时的光通量比焦距在 8~48mm 间的 F1.0 高。

计算机在实验室中已经普及，因此可以改造、共用或购买。计算机的选择通常由所使用的软件决定。有 32M 内存的苹果机或奔腾机，对 NIH-Image 和 Scion-Image 软件应用就足够了。

可以选用的凝胶分析软件很多，诸如 Macintosh 的 NIH-Image 和 PC 机 Windows 平台上的 Scion-Image，都为赠送软件，可以免费使用。在 NIM-Image 的站点，还可以得到更新的版本。Image 的帮助网页是 http://scrc.dcrt.nih.gov/imaging/tutorials/gel_density/short/index.html。通过此网址可以得到其他赠送的凝胶分析软件和其他软件或系统的经销商。Scion 正在开发新的凝胶分析软件，以便和 Scion 摄像头联合使用。也可以利用新闻组，或发标题为 subscribe 的 e-mail 到 nih-image-d-request@biomed.drexel.edu，来预订分类表单。

其他很多公司的凝胶软件都能进行比 NIH-Image 强的分析，如 Media Cyber-

netics 和 Scanalytics。在这些网站上还能找到 RFLP 和 2D 分析软件。在 14.6 节中可以找到这些公司的链接。

图像的存档很简单，但可能需要某些相册软件来管理图像文件。有许多图像存档软件可以使用，其中 Cerious 的 Thumbs-Plus 较好，因为有 Macintosh 和 PC 版本，并且有 30 天的免费试用期。该程序的 Gel-Pro 模块能提供一个图像存档数据库，自动和每个实验相联系。

随着喷墨打印机价格的下调，一般实验室可以采用如 Epson600 或 Hewlett-Packard722。热升华打印机(约 500 美元)最适合少量凝胶图像的打印，因为打印成本较高(0.5 美元/张)；喷墨打印机适合预算有限的实验室。打印大量凝胶图像最好选热敏打印机(1500 美元)，因为打印速度快，成本低(0.1 美元/张)，并且格式是专为实验记录本设计的。

14.3 方法

14.3.1 EtBr 凝胶成像系统配置示例

NIH Image 和 Scion Image 是功能相同的程序,NIH 和 Scion 都有苹果机版本，Scion 还有 Windows95/98/NT 版本。NIH 带有一些凝胶数据编辑宏文件，很多可以免费使用。这两个程序的数据和处理在很大程度上可以替换。这个例子中的硬件配置是：UVP 透射仪，Cohu 4912-5010 照相机，Scion CG-7 摄像头，NIH Image 软件,Mac 9500 计算机。对于奔腾计算机,除了安装有 Scion Image 的 Gateway 2000 外，多数设备相同。

(1) 在计算机上安装图像采集卡，并和摄像机相连。关键是打开机箱，找到 PCI 插槽，插入采集卡。

(2) 在透射仪的凝胶样本上方 14~30 英寸(30~75cm)处安装摄像头，可选用 UVP#97-0063-01 和 Bogan TC-2(货号 1882)支架。在镜头上安装 46~49mm 滤光片。在图像采集时,干涉滤光片能挡住大多数紫外线和红外线,三孔滤光片对普通 EtBr 凝胶来说就够用了。但非常暗的凝胶，可能需要更好的滤光片来减弱光源的紫外线和红外线干扰。变焦镜头需配旋进滤光片(如 ChromaTech、Corion、Omĕga Optical)。Rainbow H6X8 有 46mm×0.75mm 的螺纹，许多经销商提供的滤光片带有各种尺寸的转接环。也可以选用 AAB 的 49mm 螺纹的滤光片和本地就能买到的 46~69mm 的转接环(图 14.1)。

(3) 安装和运行 NIH Image 或 Scion Image。

(4) 在主菜单中选择 Special | Start Capturing，观察计算机屏幕上的动态视频图像。我们发现大多数凝胶在动态视图里都可见，不需更复杂的功能。所以大多数摄像头都可以用。

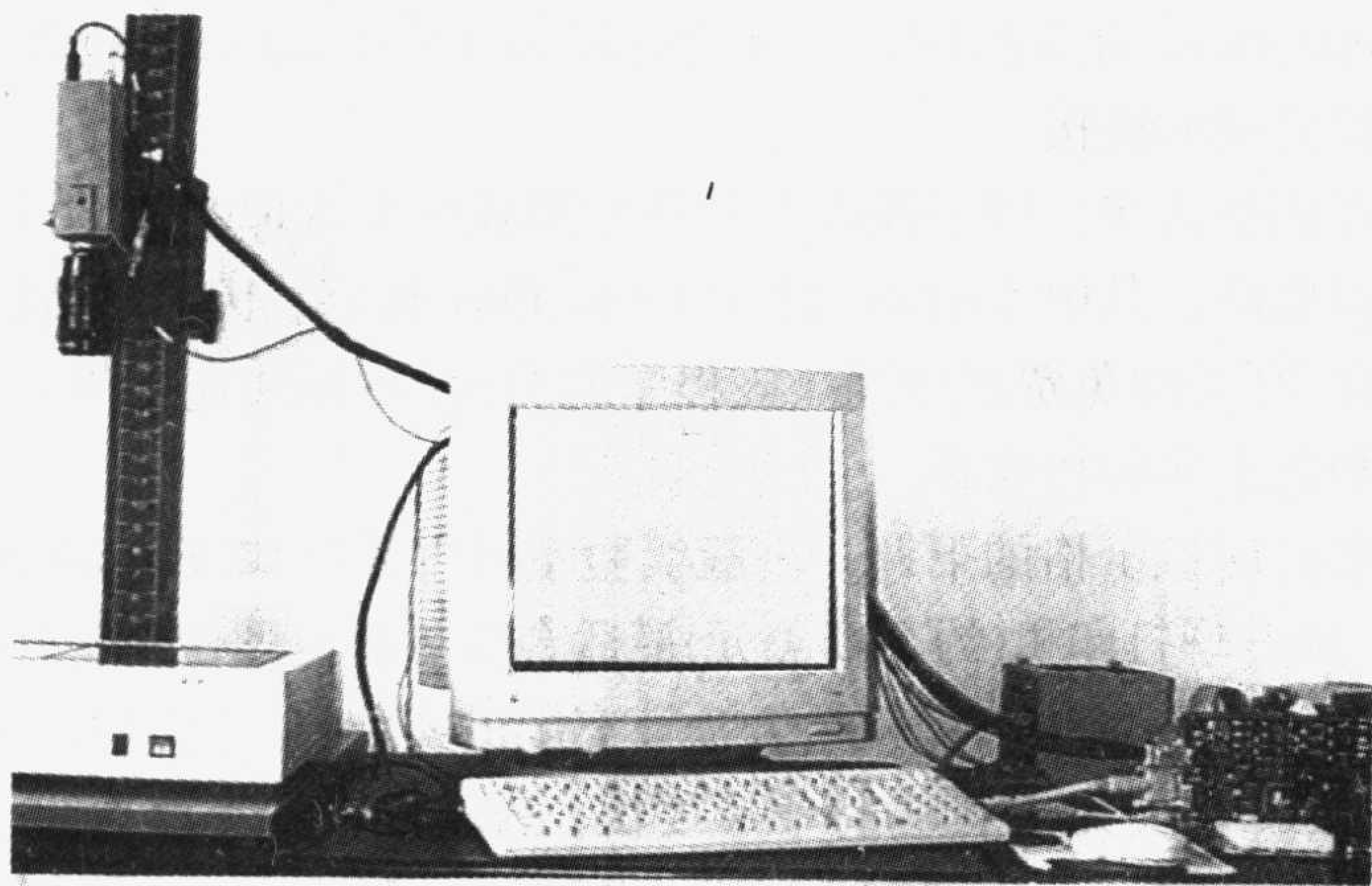


图 14.1　硬件设备图

(5) 如果在采集时看不到凝胶，用 on-chip 选项可以增加 1000 倍的灵敏度。选择 Specials | Load Macros，在 Macros 文件夹中选择 macro Video，如果用 Scion Series-7 采集卡，则选择 macro Video Series-7。

(6) 用芯片级的“on-chip”功能获取图像。如果需要最大解析度，则横向放置凝胶，让泳道水平。选择 Speicals | Continuous Integration On-chip(图 14.2)。这样将获得 4 幅图像叠加在一起的图像。把鼠标移动到图的下部，在这个区域里点击并按住左键，可以增加整合时间；在图像的上部则减少整合时间。如果采用 Scion Series-7 摄像头，选择 Continuous Integration On-chip Hi-Res，可获得 1536×1152 CCIR 解析度的超大图像。还可以选择 Special | Multi-Frame Operations 项，这是为了使用方便，保存最新整合时间所用。

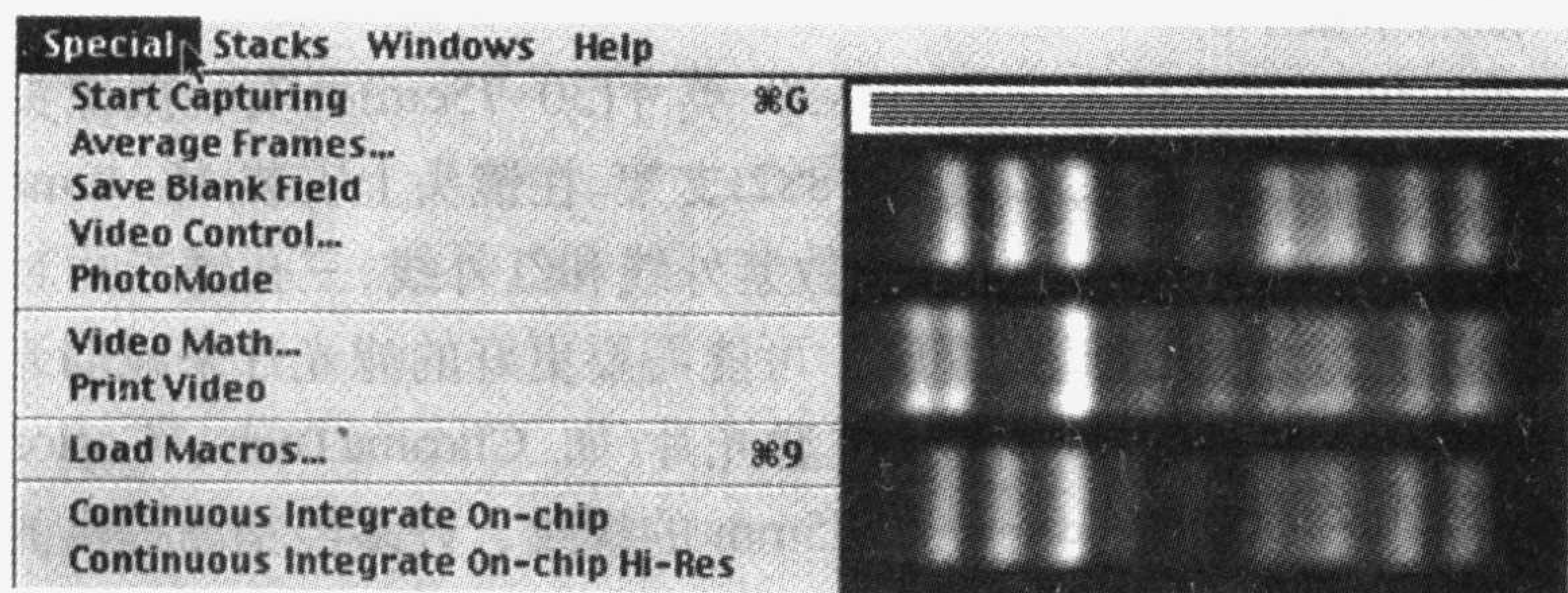


图 14.2　整合获取凝胶图像的菜单

(7) 拍摄整合图像的快捷键：可以按 Mac 机上的 period(·)键，或 PC 机上的 Esc 键。如要保存图像，选择 File | Save。保存图像时应该采用统一的记录方式，如用户名-年月日-凝胶号或年月日-实验号。可以用 Cerious 的 Thumbs Plus 程序或其他图像数据库程序来管理图像文件。一些数据库程序能加上图层，以便加上文

字标注。

(8) 光密度校准：每一像素都有在 0~256 范围内的一个亮度值。需用一套光密度标准来使 OD 值能和亮度值相互转换。选择 Scion Image 或 NIH Image 里的 Analyze 和 Calibrate。在图像窗口中，用选取工具，从光密度标准直接获得样本值。然后，选 Calibrate，键入已知的 OD 值。就可以得到数据点的对数曲线。由于图像背景为黑色，带为白色，校准前应该先翻转。可选菜单中的 Edit | Invert 项。

(9) 启用凝胶分析的宏 2(图 14.3)。

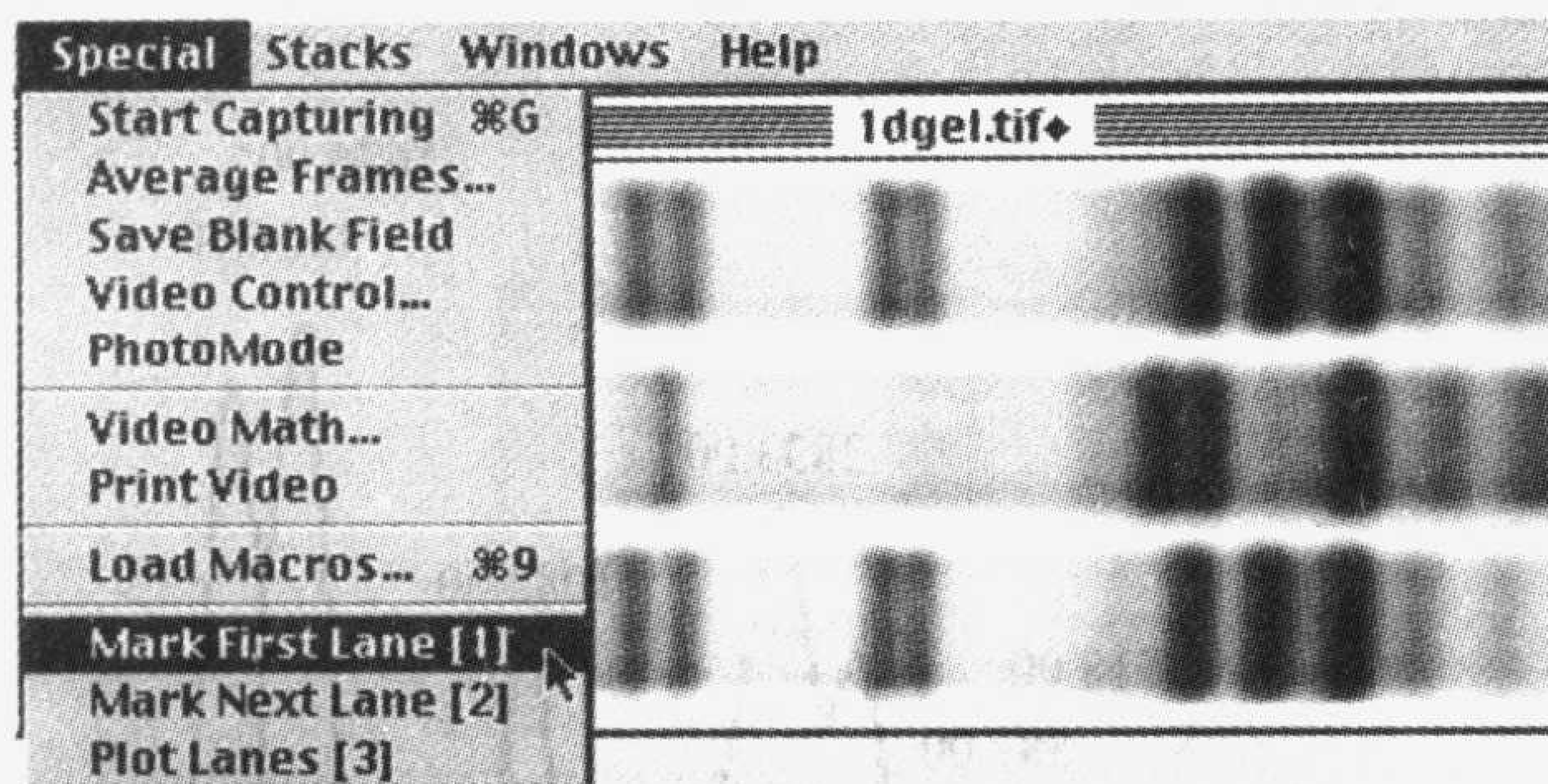


图 14.3　对凝胶反转操作分析的命令

(10) 对感兴趣的泳道，用鼠标框住泳道并加亮。如果凝胶超出屏幕宽度，可将显示器的分辨率调到最大。如果遇到凝胶泳道弯曲，可用一窄带框住。用宏命令 Plot Lanes 观察绘图曲线。NIH 程序仅能给显示部分图像加亮，所以大的凝胶需要显示器的解析度配合(图 14.4)。

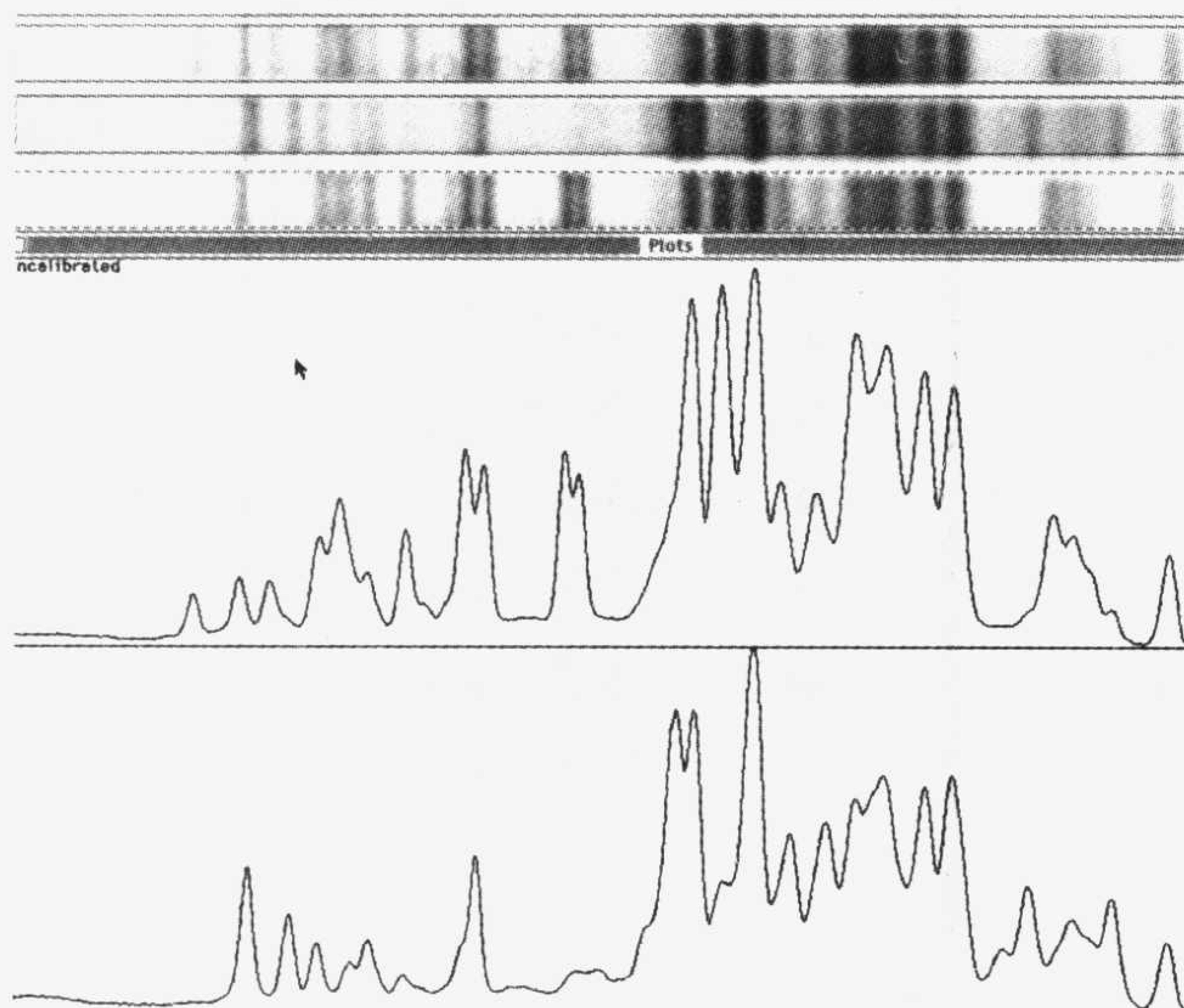


图 14.4　凝胶曲线图

(11) 用绘图工具 line 绘制基线和垂线，使每一个峰限定在一个封闭的区域内，如图 14.4 所示。按住 shift 键可确保所画的线垂直，从空泳道上取一个区域作为背景，并叠加到工作区域。

(12) 用 wand 工具连续点击每个峰内部，测定峰面积。

(13) 用 text | reverse 命令，则自动计算出选中的峰面积值(图 14.5)。(PC 机上也可以用 Scroll Lock。)峰面积也可以以表格的形式存储、显示(show results，图 14.6)、打印(print)或输出(export)。组合窗口见图 14.7。对 PostScript 打印机，为了便于印刷，PC 机的 Scion Image 可以在其网址上查询到最新的打印机清单。

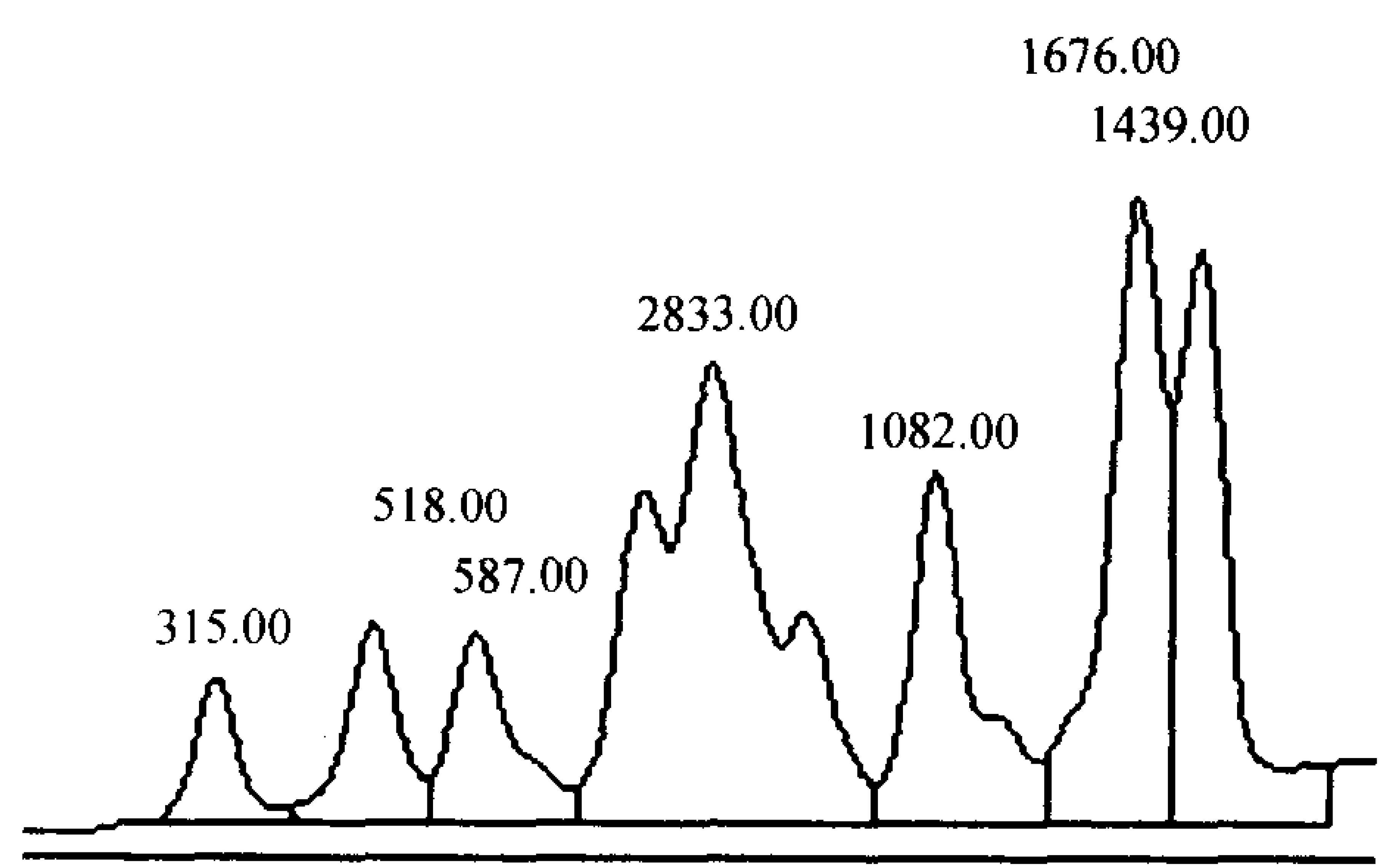


图 14.5 整合的强度测量

Results

	Area
1.	315.00
2.	518.00
3.	587.00
4.	2833.00
5.	1082.00
6.	1676.00
7.	1439.00

图 14.6 文本格式的结果

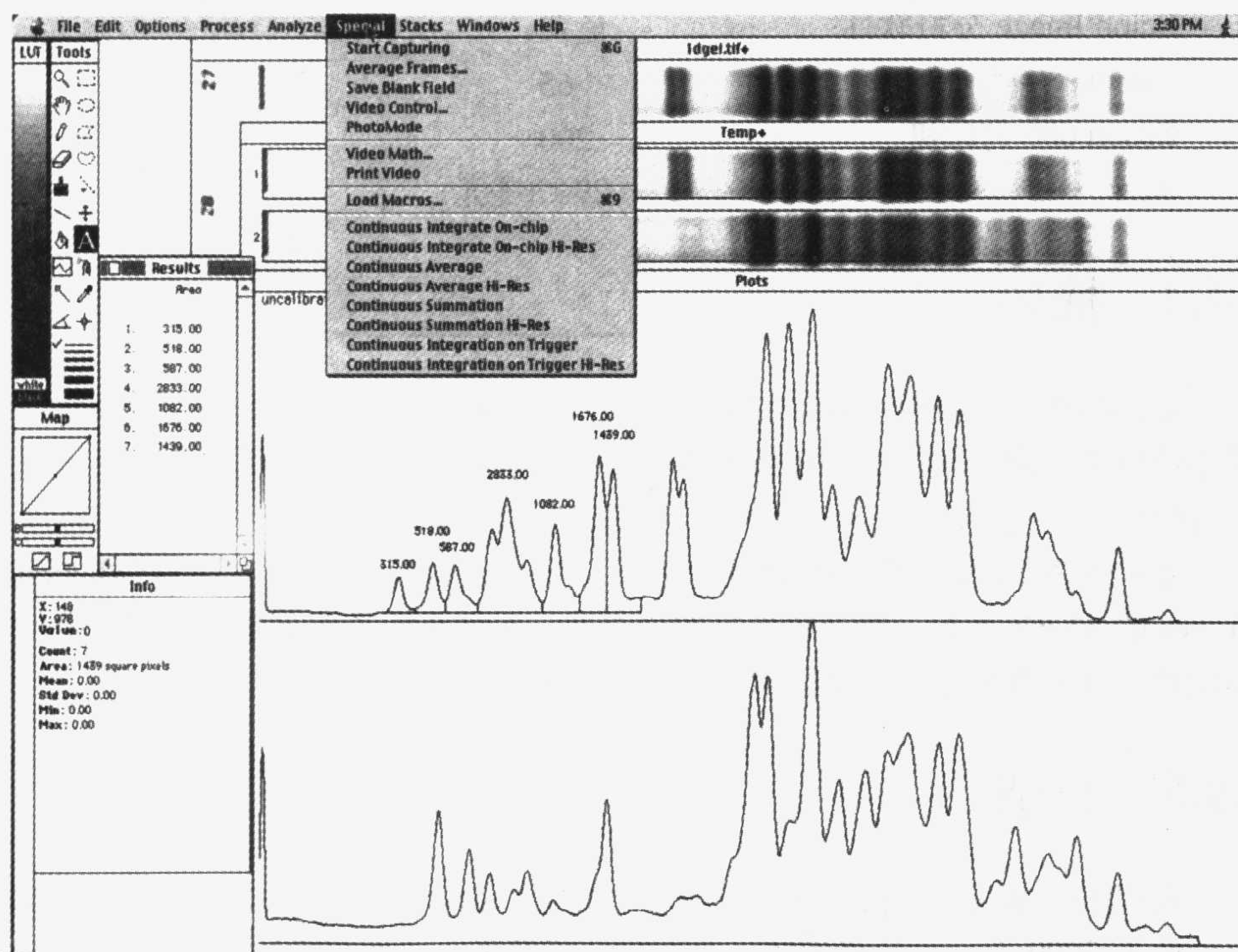


图 14.7　多个窗口的最后总结

14.3.2　Marburg 宏选项

可以利用 Marburg 宏(http://www.chemie.uni-marburg.de/~becker/image.html)修改现有的 NIH Image 宏。这些宏可以用于垂直板凝胶并能自动查询泳道、直接登录 NIH Image 网址进行分析等。

14.3.3　软件更新

从经销商那里可以获得比 NIH Image 更先进的软件包用于工作站的更新。这些软件包具有更强的功能，如自动获得泳道，在弯曲和倾斜的泳道上进行分析，以及使用解析度大于 2^8 的高清晰度照相机，见 14.6 节。

14.3.4　设备清单(不包括透射仪和计算机)

Scion GMS 300	1700
Rainbow 8~48mm	330
Omega EtBr 滤光片	250
Bogan TC-2(#1882)支架	400

Scion Image 分析软件	免费
Cerious 存档软件	65
Epson 600 打印机	200
总计	2945(美元)

14.4 摘要

本章提供了从凝胶底片到数字图像存储和分析的过程，成本低、效率高。以最小的投资实现简单的数字图像的获取和分析。数字格式的图像容易存储、再现和分析、打印或直接用 e-mail 发送。用高解析度的相机可以得到更好的清晰度，并且数码相机可以拍到更宽的动态范围。用高敏感性染料，如来自 Molecular Probes 的 SYBR 系列，能获得更好的清晰度。本文介绍了数字凝胶分析系统的基本配置，如果资金允许，可以对设备进行升级。

14.5 总结

利用实验室现有设备，购买一些补充设备，加上少于 100 美元的软件，就可以以低于 3000 美元的价格组装一台凝胶分析系统，若费用增加到 15 000 美元，就可以购买更先进的分析软件，使用更方便。

14.6 装置和软件目录

AAB，http://www.aabi.com，714-870-0290

Alpha Video，http://www.alphavideo.com，800-388-0008，612-896-9898

Bio-Rad，http://bio-rad.com/27355.html，510-741-1000

Cerious，http://www.cerious.com，704-529-0200

Chroma Tech，http://www.chroma.com，800-824-7662，802-257-1800

Cohu 4910，http://www.cohu.com，619-277-6700

Fotodyne，http://www.fotodyne.Com，800-362-3686，414-369-7000

Kodak Scientific，http://www.kodak.com/go/scientific

Media Cybernetics，http://www.mediacy.com，800-992-4256，301-495-3305

Molecular Probes，http://www.probes.com，541-465-8300

NIH Image，http://rsb.info.nih.gov/nih-image

Omega Optical，http://www.omegafilters.com，802-254-2690

Scanalytics，http://www.iplab.com，703-208-2230

Science-Intl，http://www.science-intl.com，301-631-0157

Scion Corporation, http://www.scioncorp.com, 301-695-7870
Ultra-Lum, http://www.ultralum.com, 562-529-5959
UVP, http://www.uvp.com, 800-452-6788 或 909-946-3197

致谢

衷心感谢 Scion 公司 Tod Weinberg 和 Tom Morton 的有益讨论。

(李树伟 译)

第三部分　网络信息资源

15 供临床遗传学者和分子遗传学者使用的计算机资源

Yuval Yaron 和 Avi Orr-Urtreger

15.1 引言

在过去的十几年间，遗传学已成为一门迅速发展的科学，大量日益增长的知识，使得用常规的方法(如阅读期刊和科学书籍)实际上已不可能跟踪新近的发现。这样，计算机资源对临床遗传学者和分子遗传学者便十分重要，本章的目的是为读者提供这种资源的实例。

15.2 互联网(万维网)上的分子资源

人类基因组计划(HGP)是1990年启动的一项国际合作，其目标是发现人类基因组中的50 000~100 000个基因，并使其可用于进一步的生物学研究中。所以，HGP的一个目标就是开发分析算法，集成遗传学数据库(生物信息学)，以管理和分析基因组数据，通过互联网，保健人员和研究者可以获得这些资源。大学、研究中心和卫生组织也创建了大量的数据库，可以很容易地免费查找，有一些需要注册，并只供保健人员使用，有一些注册时需要交费。在互联网上查找是有益的，因为许多数据库都互相链接，可以进入许多其他的相关站点。

15.2.1 在线人类孟德尔遗传(OMIM™)

这是一个人类基因和遗传性疾病的目录，由Johns Hopkins大学的Victor A. McKusick和其他人编辑并建立索引，本工作启动于20世纪60年代的早期，和人X连锁的性状目录一起启动。在发行了大量的印刷版本后，由美国国家生物技术信息中心(NCBI)将其开发成电子版提供给万维网(World Wide Web)。数据库含有许多出版资源、参考文献的摘要信息，并与其他互联网资源建立了链接，如Entrez：NCBI的MEDLINE和GenBank检索系统、在线动物孟德尔遗传(OMIA)、Cardiff人类基因突变数据库(HGMD)；MitoMap：Emory大学线粒体基因组数据库和其他的数据库。现在可以在万维网上得到这个有价值的资源，网址为：`http://www.ncbi.nlm.nih.gov/Omim`。

在搜索 OMIM 网页时，在适当的窗口输入关键词，关键词可包括综合名称、基因名称或指定名称，或某些临床特征。如图 15.1 所示，我们搜索“骨发育异常”，在结果网页中提供满足在搜索 OMIM 网页中输入条件的所有 OMIM 数据库列表。在所举的例子中，发现了 201 个条目，在这些条目中有 50 个被列出(图 15.2)。点击一个选项将打开某一特定的条目，本条目包括一个摘要，摘要中有临床特征和分子信息，并与其他的数据库链接，如基因组数据库(GDB)(图 15.3)。

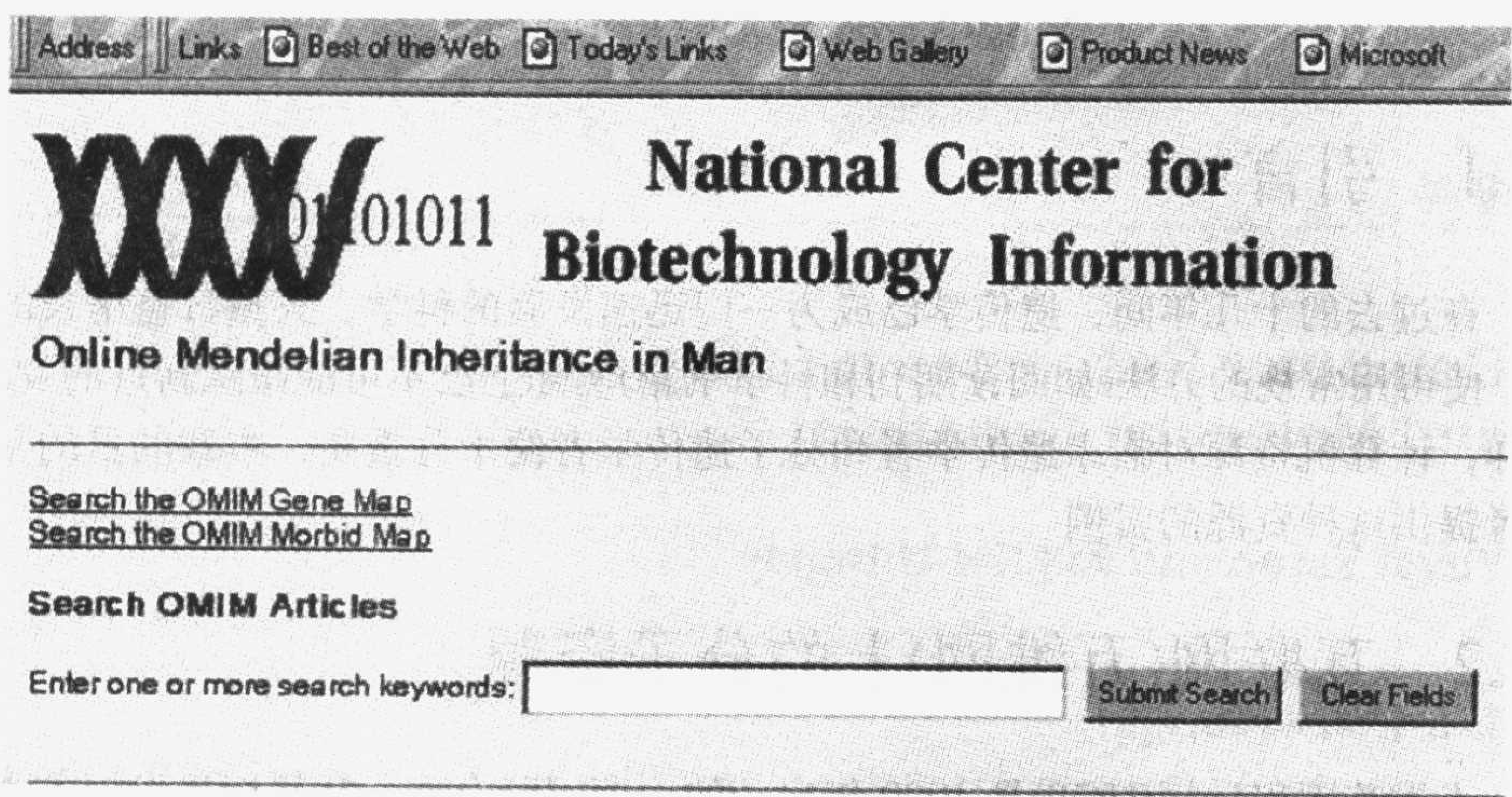


图 15.1　在线人类孟德尔遗传(OMIM)——搜索 OMIM 页面

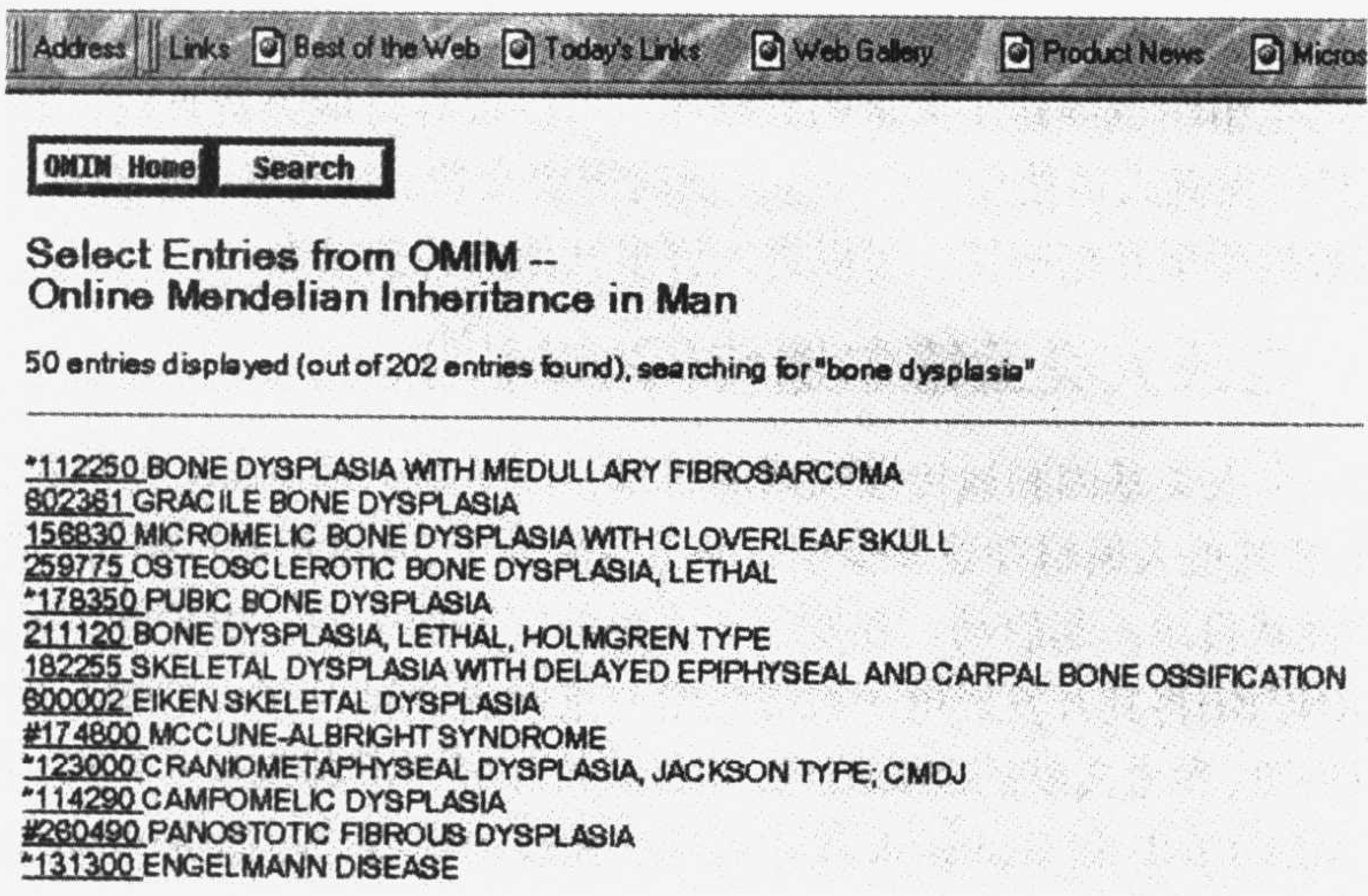


图 15.2　在线人类孟德尔遗传(OMIM)——搜索结果网页，包括所有匹配的结果

#174800 MCCUNE-ALBRIGHT SYNDROME

Alternative titles; symbols

MAS
ALBRIGHT SYNDROME
POLYOSTOTIC FIBROUS DYSPLASIA; PFD; POFD

TABLE OF CONTENTS

- DESCRIPTION
- CLINICAL FEATURES
- INHERITANCE
- MOLECULAR GENETICS
- HISTORY
- REFERENCES
- SEE ALSO
- CONTRIBUTORS
- CREATION DATE
- EDIT HISTORY
- MINI-MIM
- CLINICAL SYNOPSIS

Database Links

MEDLINE | Protein | UniGene | HGMD | Gene Map | GDB

Gene Map Locus: 20q13.2

图 15.3　在线人类孟德尔遗传(OMIM)——特定的条目、标题、别名、内容表、数据库链接、实验和参考文献

15.2.2　Entrez

本资源提供分子生物学资料和从 NCBI 集成数据库的文献引用情况，包括：来自 GenBank、EMBL 和 DDBJ 的 DNA 序列；来自 SwissProt、PIR、PRF 和 PDB 的蛋白质序列；从 DNA 序列数据库翻译的蛋白质序列；基因组和染色体作图资料；来源于 PDB 的蛋白质三维结构，并加入到 NCBI 的分子模建数据库(MMDB)。此外，通过国家医学图书馆的 MEDLINE 和 Pre MEDLINE 数据库，可以得到引用近 900 万篇生物医学文章的文献数据库(PubMed)。其网址为：`http://www.ncbi.nlm.nih.gov/Entrez`。

15.2.3　美国国家生物技术信息中心

本网站提供大量与基因序列数据库的链接，如 Entrez、Gene Map of the Human Genome、dbEST(表达序列标签数据库)，也提供与其他 NCBI 资源的链接，如由国家医学图书馆(NLM)提供的 PubMed(免费 MEDLINE)，其网址为：`http://www.ncbi.nlm.nih.gov/`。

15.2.4　Cardiff 人类基因突变数据库

本数据库包括大量已知的人类遗传性疾病的基因突变，包含人类核基因编码区内的各种类型的突变，已知这些突变引起人类的遗传性疾病，本数据库不包括

没有明显表型结果的多态性。本数据库由位于 Cardiff 的威尔士大学医学院的医学遗传学研究所的 D. N. Cooper、E. V. Ball、P. D. Stenson、M. Krawczak 等人维护。其网址为：http://www.uwcm.ac.uk/uwcm/mg/hgmd0.html。

15.2.5 基因卡(GeneCard): 人类基因、蛋白质和疾病数据库

基因卡数据库集成人类基因及其产物的信息，储存在主数据库中，资料是从以下数据库中提取的：GDB，包括基因和其他基因组特性的信息；MGD(鼠基因组数据库)，包括实验小白鼠的实验遗传学信息，包括标记物、哺乳动物同源性、探针和克隆；OMIM，人类基因和遗传性疾病目录；SwissProt，包括蛋白质、其序列和细胞内功能的信息；HGMD，有关基因中致病突变的信息资源；互联网医师指导，网上服务，刊登有关生物医学研究及其应用的新闻；基因卡，基因、标记物和表型的目录，与主要的数据资源建立了链接。大多数的信息由编写的用来在其他数据库中搜索有关这些基因信息的脚本自动加入。得到资料后，进行关联性分析、写出摘要并转化为数据库的活动条目。基因卡是 Weizmann 研究所基因组中心和 Weizmann 研究所生物信息部的一个项目。基因卡的概念、脚本和网页界面是由 Michael Rebhan 和 Jaime Prilusky 与 Vered Chalifa-Caspi、Marilyn Safran、Liora Yaar 和 Doron Lancet 合作开发的，其网址为：http://bioinformatics.weizmann.ac.il/cards/(图 15.4)。

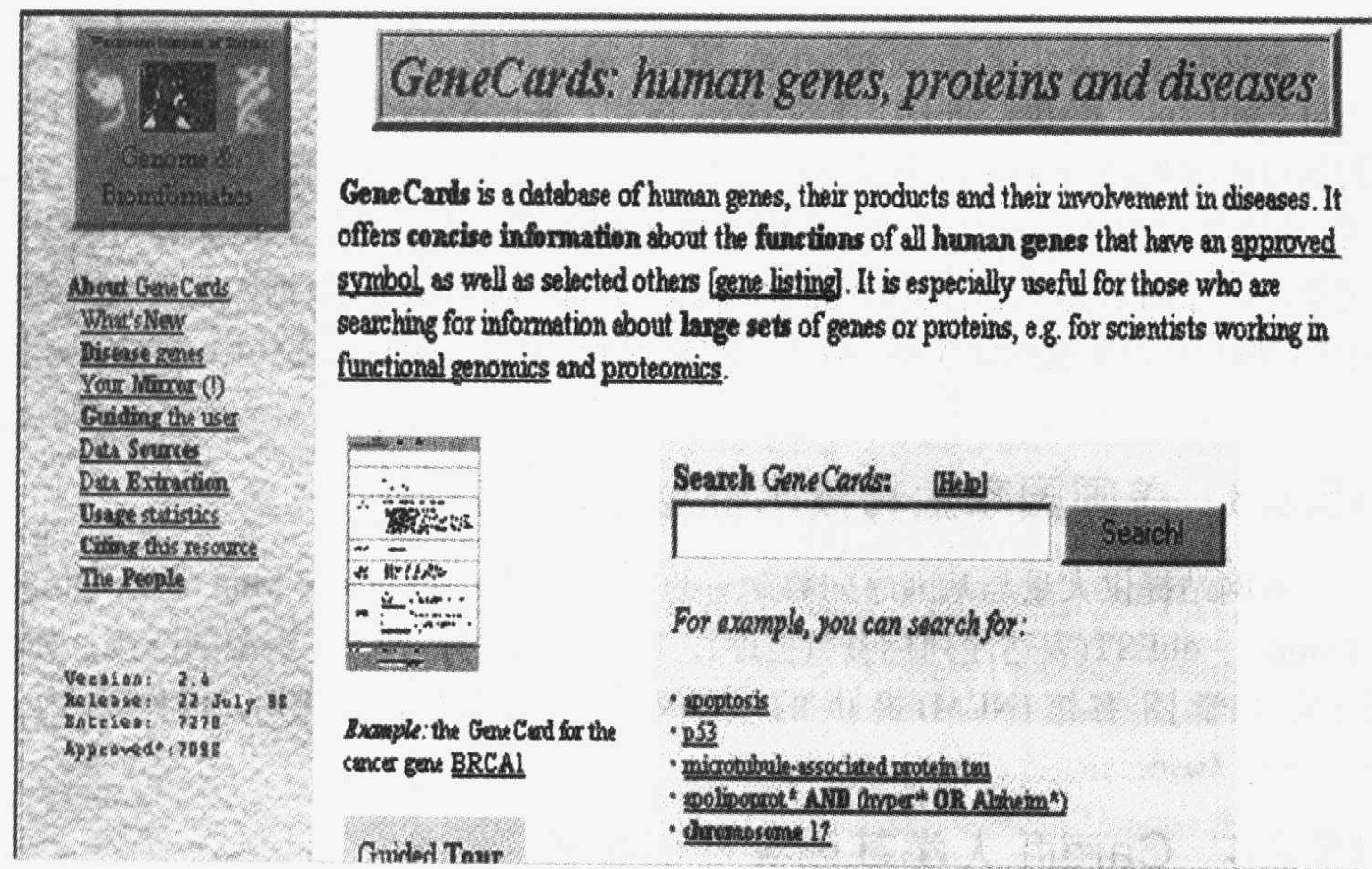


图 15.4 基因卡：人类基因、蛋白质和疾病数据库

15.3 互联网上的临床资源

15.3.1 供遗传学专业人员用的信息——堪萨斯大学医学中心

本资源提供定期更新的供遗传学专业人员使用的信息，并与临床和研究资源建立了链接，如遗传学会、各种遗传学疾病的专家组、临床遗传学数据库(OMIM, GeneTests)和遗传计算机资源。其网址为：`http://www.kumc.edu/gec/geneinfo.html`(图 15.5)。

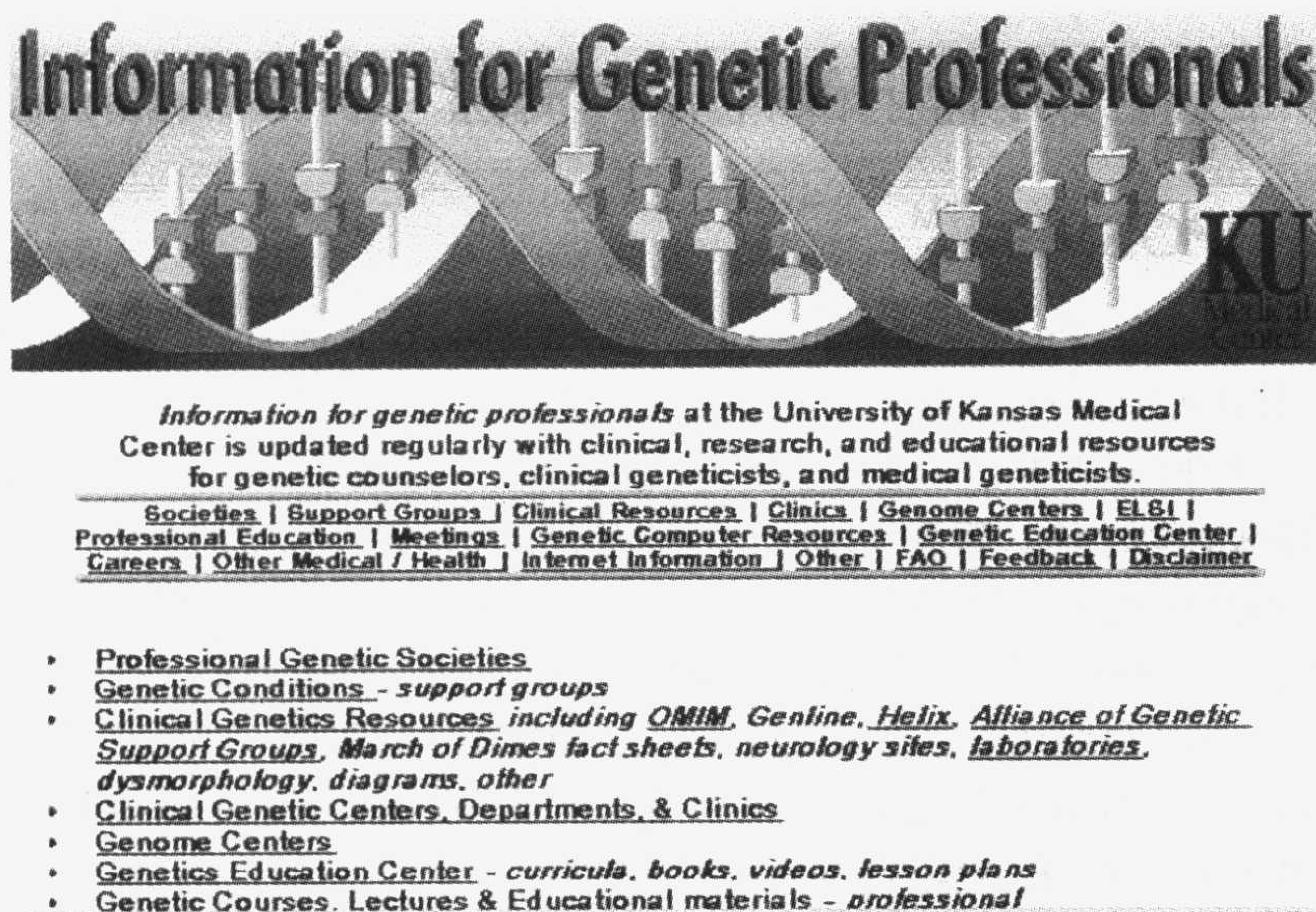


图 15.5 供遗传学专业人员用的信息——堪萨斯大学医学中心

15.3.2 GeneTests™

GeneTests™ 是一个对遗传性疾病进行疾病特异性诊断测试和/或研究测试的实验室目录，含有 300 多个实验室的列表，测试 550 多种遗传性疾病。本资源由国家医学图书馆提供经费，由儿童医院、地区中心和华盛顿大学医学院维护。本资源是免费的，但只限于保健人员使用，必须注册，凭密码进入。其网址为：`http://www.genetests.org`，通过 e-mail、电话或传真也可以进入。

15.3.3 Reprotox®

本数据库提供最新的有关处方药、非处方药和毒品，以及工业和环境化学品对生殖的影响的信息。每个条目都有摘要，提供所查询药对人、动物和体外研究

的资料，信息覆盖人类生殖的所有方面，包括生育力、雄性暴露(male exposure)和哺乳，特别注重胚胎和婴儿暴露的结果。搜索按钮打开一个搜索窗口，可以根据关键词进行搜索。如商标名、属名、街道名、化学品和其他环境暴露。搜索结果网页提供所有可能条目的列表，这些条目都与所搜索的关键词有关联性，以可信度程度来分类，选择任何一个选项都可以提供条目的信息摘要。

每年交预定费后，由生殖毒理学中心(RTC)提供服务，该中心位于哥伦比亚医院女性医疗中心，地址为：2440 M Street，NW，Suite 217，Washington，DC 20037-1404。每年交费后，也可以在 http://reprotox.com 得到该数据库，也可以从磁盘或光盘上得到，有 DOS 或 Windows 版本。

15.4 磁盘或光盘上的计算机资源

15.4.1 POSSUM™

POSSUM 是“Pictures of Standard Syndromes and Undiagnosed Malformations”[见 15.5 注(1)]的首字母缩略词，它是一个诊断临床综合征的工具软件，现在可以从光盘上得到(5.0 版本)，可在标准的 PC 上运行，不需要特殊的激光盘(以前的版本需要特殊的激光盘)，新版本也从 OSSUM(有关骨骼发育异常的系统，链接几千张 X 光片)合并信息。用户提供一张患者症状的列表，软件提供一张可能综合征的列表。它含有 3000 种综合征，2000 种患者表现的出生缺陷模式。数据库包括大量的图片，包括放射线学和临床特征，并与 OMIM 自动链接。该开发是由澳大利亚墨尔本的 Murdoch 出生缺陷研究所和皇家儿童医院的维多利亚临床遗传学者服务处的 David Danks 和 Agnes Bankier 所领导的。

15.4.2 伦敦畸形学数据库和神经遗传学数据库(LDDB，LNDB™)

这些数据库是由牛津医学数据库建立的，作者是儿童健康研究所的临床遗传学和胎儿医学母亲保健部的 Robin Winter 和 Michael Baraitser，其地址为：30 Guilford Street，London，WClN IEH[见 15.5 注(2)]。该数据库为临床遗传学者提供先天性畸形和神经遗传学综合征临床诊断的一个工具，这些数据库是从 1000 多种期刊中编辑的，有关于综合征的许多参考文献。畸形学数据库包括 2750 多种单基因疾病、偶尔发生的疾病和那些由环境因素引起的疾病，染色体异常未包括在内。神经遗传学数据库包含 2500 多种综合征的信息，包括中枢和外周神经系统。附带的磁盘为照片库，含有影像。从 1996 年起有了 Windows 版本，开始的窗口可以在畸形学数据库和神经遗传学数据库之间选择，在每个数据库中，主窗口含有数据库中所有综合征的列表。如果知道一种特定综合征的名称，则可以直接搜

索，然而，在临床设置中，常常不知道正确的诊断，这样，可以通过关键词搜索或基于特征的搜索选项来查询数据库。关键词搜索选项通过在标题、摘要或参考文献中出现的关键词、通过染色体定位、通过 OMIM 号或通过遗传方式进行搜索。基于特征的搜索选项是本软件最出色的特征，在适当的窗口中输入不同的规则，像选择临床特征的分级列表一样。特征列表的开始是常规的系统分类，之后是二级的无显著特点的临床特征，最后是更低水平的特定临床特征。如果在分类列表中指定了特定的规则，则发现特征选项可以根据特定的规则进行目标追踪，可以根据所有指定的规则或其中的几个来进行搜索，而且，某些规则可以被标记为必选项，以保证所发现的所有综合征都满足指定的规则。一旦定位了一种综合征，则可进入该综合征的细节，包括原文摘要、特征列表和参考文献。其他综合征细节包括细胞生成的位置、OMIM 号、综合征列表和综合征遗传性的细节。在窗口上有一套“拇指甲”，表示在照片库中匹配的影像。

15.4.3 人类细胞发生数据库

人类细胞发生数据库是由苏黎世的医学遗传学研究所的 Albert Schinzel 编辑的，由牛津大学出版社以磁盘的形式出版，作为牛津医学数据库系列的一部分，总编为 Michael Baraitser 和 Robin Winter 教授。该数据库能够搜索与 1000 多种染色体异常有关的临床和细胞发生资料，并通过搜索临床特征列表可以选择染色体异常。该数据库最重要的一个特点是它能够添加自己的数据而得到扩充。

15.5 注

(1) POSSUM：Anne Cronin 皇家儿童医院 Murdoch 研究所，
Flemington Road，Parkville，Victoria Australia，3052，Tel: +61 3 9345 5045，e-mail: cronin@cryptic.rch.unimelb.edu.au。

(2) 人类细胞发生数据库，LDDB 和 LNDB：Janet Caldwell 或 Rachel Rains，电子出版社，牛津大学出版社，Great Clarendon Street，Oxford OX2 6DP，UK.，Tel：(01865) 267979，e-mail：ep.info@oup.co.uk。

15.6 引用资源的网址

(1) 在线人类孟德尔遗传(OMIM)，http://www.ncbi.nlm.nih.gov/Omim。

(2) Entrez，http://www.ncbi.nlm.nih.gov/Entrez。

(3) 国家生物技术信息中心(NCBI)，http://www.ncbi.nlm.nih.gov/。

(4) Cardiff 人类基因突变数据库，http://www.uwcm.ac.uk/uwcm/mg/

hgmd0.html。

(5) GeneCards:人类基因、蛋白质和疾病百科全书，http://bioinformatics.weizmann.ac.il/cards/。

(6) 供遗传学专业人员用的信息——堪萨斯大学医学中心，http://www.kumc.edu/gec/geneinfo.html。

(7) GeneTests，http://www.genetests.org/。

(8) REPROTOX，http://reprotox.com。

(李慎涛　译)

16 NCBI 网页上的公用工具和资源

Jack P. Jenuth

16.1 引言

美国国家生物技术信息中心(NCBI)建于 1988 年 11 月，建在美国的国家医学图书馆(NLM)内。之所以选择国家医学图书馆，是因为它有建立和维护生物医学数据库的经验，而且作为国家健康研究所(NIH)的一个单位，它能够在计算分子生物学方面建立一个内部的研究项目。NCBI 的任务是开发新的信息技术，来帮助理解控制健康和疾病的基本分子过程和遗传过程。从 NCBI 网页站点(`http://www.ncbi.nlm.nih.gov/`)上便可以知道，它主要有以下四项任务：

(1) 建立储存和分析有关分子生物学、生物化学和遗传学知识的自动系统。

(2) 进行计算机信息处理高级方法的研究，以及分析生物学上重要分子的结构和功能。

(3) 使生物技术研究人员和医学界人士更容易地使用数据库和软件。

(4) 共同合作来收集世界范围内的生物技术信息。

在过去的 10 多年间，这些任务的结果被科学界以多种方式广泛利用，这包括通过磁性介质或 CD-ROM，以及通过使用诸如 ftp、gopher、e-mail 和万维网等方法来发行数据库和软件。本章叙述用户通过万维网可以得到的工具和资源。

16.2 GenBank

16.2.1 GenBank 介绍

GenBank[1]是一个 DNA 序列数据库，建于 1980 年代的早期。从 1980 年代后半期至 1992 年 10 月，该数据库由 Intelligenetics 维护，之后，由 NCBI 负责维护。在这段时间内，GenBank 在核苷酸数量和条目数量两方面都呈对数增长，如图 16.1 所示。到 1998 年为止，在 2 209 000 个序列记录中大约有 1 500 000 000 个碱基。

经过训练的具有大学水平生物学经验的检索员从科学文献中将数据记录录入

到 GenBank 中，由作者直接提交数据使得信息得到扩大。国际核苷酸序列数据库协会——包括欧洲分子生物学实验室(EMBL)和日本 DNA 数据库(DDBJ)的会员可以进行数据协作和交换。这些组织每天都交换数据。与国家农业图书馆和美国专利商标局联合，能够加入植物和专利序列的数据。

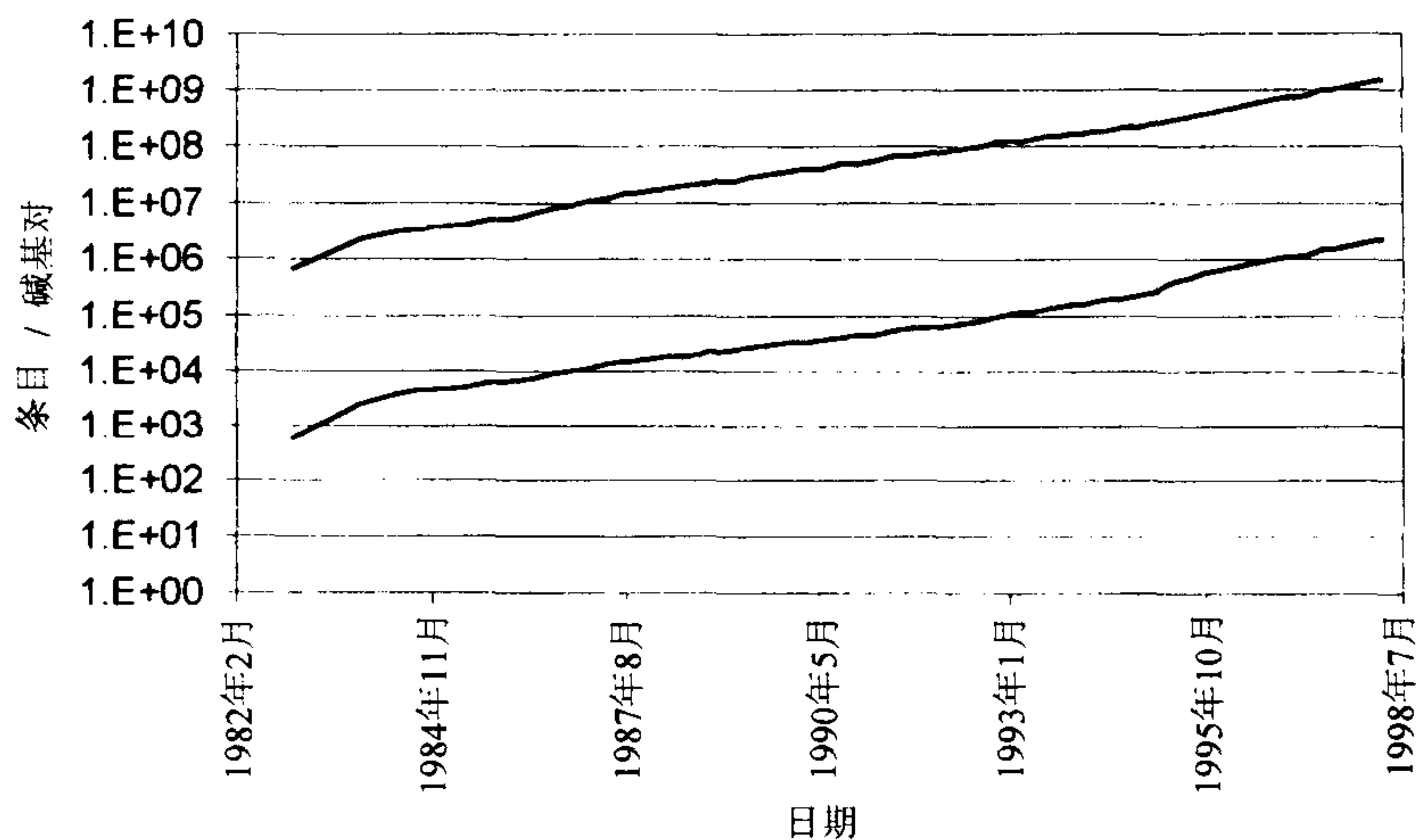


图 16.1　GenBank 容量的增长

上面的线表示核苷酸的数目，下面的线表示 GenBank 中条目的数目

可以通过几种不同的方法进入 GenBank 数据库，如通过使用列表服务器的 e-mail、ftp 和通过万维网(使用大多数网页浏览器)。网页界面允许你提交遗传信息，可以根据核苷酸序列和条目内的注解进行各种查询。

16.2.2　数据提交和校正

通过当地的一个叫作 Sequin 的程序或通过一种基于在线的网页服务(叫作 BankIt)可以将数据提交到 GenBank。Sequin 是一个基于 PC 的程序，它可以使用户注解一个序列、进行一定的分析并将其提交给 GenBank。这个程序有各种平台的版本，包括 Mac、Windows 和 Unix。当使用 Sequin 时，直接提交的输出文件可以通过 e-mail 送到 GenBank，网址为：`gb-sb@ncbi.nlm.nih.gov`，或将提交的文件拷贝到软盘上，邮寄到 GenBank 提交处。基于网页的形式称为 BankIt，从 NCBI 的主页上(`http://www.ncbi.nlm.nih.gov/BankIt/index.html`)点击“BankIt”便可以发现。从这里，你可以进入一个序列或校正一个序列。在 BankIt 的主页上，可以见到逐步的说明，指导你进入新的序列或对已存在的序列进行更新。

16.2.3 搜索 GenBank

16.2.3.1 BLAST

NCBI 的研究人员开发了一个称为 Basic Local Alignment Search Tool(BLAST)的程序，来快速地将一个氨基酸或核苷酸的查询序列与序列数据库进行比较。在一个查询序列和一个数据库之间，BLAST 并不一定发现最好的命中(hit)，它可能发现不了匹配，这是因为它使用这样一种策略(大约法)：希望发现最多的匹配，但为了获得速度而牺牲了完整的灵敏度。BLAST 的最新版本使用了一种新的算法，搜索速度较首次发行的版本要快，可以产生空位比对(gapped alignment)(头一版本无此功能)，在 psi-BLAST 版本中灵敏度得到提高。

在万维网上,进入NCBI的主页,点击BLAST图标(`http://www.ncbi.nlm.nih.gov/BLAST/`)便可以得到 BLAST。从这里你可以选择使用几种不同版本的 BLAST 对数据库进行搜索：

(1) BLAST：从这些主页进行的搜索使用 BLAST 1.x 版本[2]。这些搜索比 BLAST 2.0 要慢，在报告的命中比对中没有空位(gap)。

(2) BLAST 2.0：是 BLAST 的一个新版本，明显提高了搜索的速度[3]。而且，BLAST 2.0 能够在报告的命中比对中引入空位。

(3) Psi-BLAST：是位置特异性迭代 BLAST 的缩写词，它使用来自于一个 BLAST 的任何有意义的比对信息来构建一个位置特异性的评分矩阵(score matrix)，在下一轮的数据库搜索中，用其代替查询的序列。使用这种方法，能够提高搜索的灵敏度，只要从首次 BLAST 搜索得到的比对有效，便可以找出远缘相关的序列。

可进行的搜索类型见表 16.1。数据库来源于大量的资料，不必直接对应于目前在 GenBank 中所列举的数据库。每个数据库的名称及对其的简述见表 16.2。

表 16.1 使用 BLAST 类程序进行的有效查询

程序	查询序列	数据库类型	用于比较的程序
BLAST/BLAST2	核苷酸	核苷酸	blastn
	核苷酸(翻译的)	蛋白质	blastx
	蛋白质	核苷酸(翻译的)	tblastn
	核苷酸(翻译的)	核苷酸(翻译的)	tblastx
	蛋白质	蛋白质	blastp
Psi-BLAST	蛋白质	蛋白质	blastp

表 16.2 NCBI 上的 BLAST 数据库

数据库	简　述
肽序列	
nr	所有非冗余 GenBank CDS 翻译+PDB+SwissProt+PIR
month	在最近 30 天内发表的所有新的或改进的 GenBank CDS 翻译+PDB+SwissProt+PIR
swissprot	SwissProt 蛋白质序列数据库最新的主要版本
yeast	酵母(酿酒酵母)蛋白质序列
E.coli	大肠杆菌基因组 CDS 翻译
pdb	从三维结构 Brookhaven 蛋白质数据库推导出的序列
kabat[kabatpro]	Kabat 免疫学目的序列数据库
alu	由 REPBASE 中选择的 Alu 重复序列翻译而来，用来遮蔽查询序列中的 Alu 重复序列
核酸序列	
nr	所有非冗余 GenBank+EMBL+DDBJ+PDB 序列(但不含 EST、STS、GSS 或 HTGS 序列)
month	在最近 30 天内发表的所有新的或改进的 GenBank+EMBL+DDBJ+PDB 序列
dbest	GenBank+EMBL+DDBJ EST 部门的非冗余数据库
dbsts	GenBank+EMBL+DDBJ STS 部门的非冗余数据库
htgs	高产率基因组序列
yeast	酵母(酿酒酵母)核苷酸序列
E.coli	大肠杆菌基因组核苷酸序列
pdb	从三维结构推导出的序列
kabat[kabatnuc]	Kabat 免疫学目的序列数据库
vector	载体
mito	线粒体序列
alu	从 REPBASE 选择 Alu 重复序列，用来遮蔽查询序列中的 Alu 重复序列
epd	真核启动子数据库
gss	基因组概览序列，包括单向(single-pass)基因组数据、外显子捕获序列和 Alu PCR 序列

要进行一个搜索，选择 BLAST 搜索页，将你的目的序列进行剪切和粘贴、选择程序和数据库、提交。进行一个基本的 BLAST 搜索，唯一可以选择的是进行 BLAST2 的一个空位比对，以及过滤掉低复杂度区域的序列。过滤选择只是简单地将比如 polyA 尾、poly-谷氨酰胺序列、重复序列等的区域(即低复杂度的区域)替换成 Ns(核酸序列)或 Xs(肽序列)。低复杂度的区域通常产生假的高分，这只反映出组成偏差，而不反映有意义的位置对位置的比对。如果这个不

适合于你的查询，只要关掉此项选择即可。对大多数查询来说，不需要改动默认的参数。当然，用户使用高级 Blast 页的确会改动某些参数的选项。这些选项见表 16.3。

表 16.3　BLAST 程序的可选项

可选项	简　述
Expect	报告匹配的统计学上有意义的阈值。只有当值小于输入的数时才表明匹配
NCBI-gi	在输出结果中，除了基因存取号和/或基因位点名称外，还表明命中的 NCBI-gi 号。当序列更新时，gi 改变，但基因存取号永远不变
Descriptions	返回命中的最大数
Alignments	返回比对的最大数(总是小于返回的叙述数)
Graphical overview	展示与查询序列比对的数据库序列的概况。每个比对的得分被分成 5 组，分别用 5 种不同的颜色表示。在同一数据库序列中，用线条连接多个比对。将鼠标光标对准一个命中序列，将在窗口的顶部出现定义和得分，点击一个命中序列，可使用户进入相关的比对
其他高级可选项	
Cost to open gap, cost to extend gap	这些值只用于产生比对，在此不详细讨论这些可选项
Reward for nucleotide match and penalty for nucleotide mismatch	匹配：不匹配的比率决定了发现核苷酸命中的灵敏度。对更趋异的种属增加此比率。默认值为：reward = 1，penalty = −3。本参数不适合于肽数据库序列
Word size(W)	字体大小限制程序只发现那些有一段“W”核苷酸序列或氨基酸序列与查询序列 100 %同源的序列。blastn 的默认值为 11，其他程序的默认值为 3。对于核苷酸序列，降低此数可以增加灵敏度

对进行的每一次搜索，返回一个有三部分组成的结果。第一部分是对数据库序列与查询序列比对情况的图形总结。每一个比对的得分被分成五组，分别用不同的颜色表示。在同一个数据库内，用线条连接多个比对。在一个命中上点击可以将定义和得分列在顶部的窗口。在一个命中上点击可以将用户带到相关的比对。第二部分是对这些命中、登录号、期望值和/或 gi 号及得分的简述。点击得分会将你带到查询序列与命中(第三部分)的比对。你也可以点击 gi 或登录号，这样会将查询登录到 Entrez 系统中，并返回所选命中完整的 GenBank 条目。每个 GenBank 条目含有与其他数据库的链接，这些数据库与本条目交叉对照。这些链接能够使用户迅速地收集到每个命中的其他信息。可能包含的有用链接有：MEDLINE、OMIM、序列特征数据库、相关的肽或核酸条目等。

对许多物种序列不完整的基因组也可以进行 BLAST 搜索。在 NCBI 的主页上点击 Entrez，接着点击基因组就可以进入这些网页。接着出现一个多页框，从此处你可以选择 BLAST WITH UNFINISHED MICROBIAL GENOMES。除具有选择指定你要查询的微生物基因组的功能外，本网页与其他 BLAST 网页相同。

16.2.3.2 Entrez

Entrez 是一个能够使用户在许多 NCBI 维护的数据库上进行文本搜索的系统。这些数据库包括以下内容：

(1) 来自 GenBank、EMBL、DDBJ 的 DNA 序列。

(2) 来自 SwissProt、PIR、PRF、PDB 的蛋白质序列，以及来自于 DNA 序列数据库的翻译的蛋白质序列。

(3) 基因组和染色体图资料。

(4) 来自于 PDB 的三维蛋白质结构，并与 NCBI 的分子模建数据库(MMDB)结合。

(5) 来自于国家医学图书馆的 MEDLINE 和 pre-MEDLINE 数据库的 PubMed 文献数据库。

要进入 Entrez，只需点击 NCBI 主页上的 Entrez 按钮即可。从打开的网页可以搜索上述数据库之一，不同的数据库所呈现的形式是相同的。有两种方式进行搜索：自动方式，将接受指定的术语并自动与从数据库检索到的术语进行匹配；术语列表方式(list-term mode)，对每个搜索给出选择检索到的术语的可选项。一旦一个搜索术语被选中并提交，便返回一个显示命中号码的页，并出现一个对话框，可进一步进行搜索。在一个术语的后面加一个星号，Entrez 将搜索所有以此词开头的术语。星号后面单词之间有空格的短语将不包括在内，将词放在括号内可以迫使 Entrez 搜索短语，并尽量发现你输入中的逻辑编组。要进一步搜索，只需选择你要搜索的领域，键入新的术语，提交查询。返回一张相似的表格，显示新命中的号码。你可以进一步搜索，直到文章的数目较小为止，此时只需点击检索文档键即可返回命中。

当从查询页检索到文档或序列时，将呈现出每个命中的摘要。每个命中可以用几种不同的格式观看，例如，核酸序列用 GenBank、FASTA 或 ASN.1，肽用 GenPept、FASTA 和 ASN.1。也可以呈现图的形式，列出肽或核酸已知特性的位置，你可以点击任何的特性来显示进一步的细节。对 GenBank 中的每个条目，对具有明显同源性的序列以前已进行过计算，点击 nucleotide 或 protein neighbors 键便可以进入。

此外，每个 GenBank 条目都与其他数据库进行了交叉参考。点击有蓝色下划线的文本便可以直接浏览这些交叉参考文献，本特性使你能够快速地收集其他信息。

16.2.4 基因组资源

从许多原核和真核生物积累的遗传学文献正在日益增长，已有许多有用的工具通过同源性搜索来分析蛋白质的功能、建立蛋白质的进化关系、确定不同物种之间基因和蛋白质的相互关系。这些资源包括同源基因信息库和 UniGene。

16.2.4.1 同源基因信息库(COG)

通过比较在 7 个完整基因组(代表 5 个主要的系统发育谱系[4])中编码的蛋白质序列，已经计算了基因的种间同源性(在物种形成过程中，不同的种从相同的祖先基因中进化而来的基因)。在进化过程中，这些种间类似基因通常保留相同的功能。这对于描述在其他种中相关蛋白质的功能是十分有用的，例如，对细菌中广泛可见的蛋白质的分类，从而可以作为新抗生素的有用的靶。

你可以浏览 7 个完整基因组当前分析的结果，包括大肠杆菌、流感嗜血杆菌、生殖道霉形体、肺炎霉形体、蓝细菌属-集胞蓝细菌属、詹氏甲烷球菌和酿酒酵母。由已测序 DNA 编码的蛋白质分成 4 个总的类型：信息储存和加工、细胞内加工、代谢和尚未完全定性的。这些种类的每一种又进一步分成几组，在每组中包含 COG。点击功能和系统发生分析下的 List of all COGs 或 Table，可以浏览 COG，显示出每个 COG 的模式、大小和成员数。如果进一步点击 ID，伴随图形显示重叠区域、最佳命中草图和一个簇进化树，显示出每个 COG 的肽成对的比对。

要将你自己的序列与现有的 COG 进行比较，选择 Cognior 键，将显示一张网页，在此网页上，可以将你自己的序列剪切和粘贴到一个框内，并提交查询。对 COG 数据库进行一个 BLAST 搜索，可得到与查询序列单个命中比对的文本 BLAST 结果和图形显示。而且，还可得到最佳命中的草图，此草图显示你的查询与 COG(与你的序列匹配最佳)之间的关系。黑色的实线代表对称的最佳命中(BeTs)，虚彩线代表不对称的 BeTs。例如，一种酵母蛋白与一种大肠杆菌之间的绿色线表明已知的大肠杆菌蛋白是已知的酵母蛋白的最佳命中，但反过来则不正确。

16.2.4.2 UniGene

UniGene 是一个独特的人和鼠序列数据库，它是通过对重叠的 EST 进行分簇而衍生来的。UniGene 也包含在 GenBank 中发现的非冗余的 cDNA 和 CDS。GenBank 106 版有 47 000 个 UniGene 簇、14 000 个独特 cDNA 和 CDS。在这些 cDNA 中，只有 500 个没有相关的 EST。UniGene 每两个月更新一次，大约比一个新的 GenBank 版本晚一周。可以从 NCBI 的 FTP 站点的 repository/UniGene 子目录下下载文件。

要查询 UniGene，选择 UniGene：Unique Gene Sequence Collection for Human

and Mouse，下一屏含有一个 UniGene 的描述，在左上角有两个按钮，一个将带你进入鼠 UniGene 组，另一个进入人类簇。人类和鼠 UniGene 组的查询窗口完全相同。你可以在提供的框内键入一个搜索术语，便可搜索 UniGene 的注解。有效的查询包括搜索术语(如磷酸酯酶)和 GenBank 的登录号。此外，还提供了许多@ 功能，以用于特定的目的，点击 query tips 便可见到对其的叙述。点击染色体号，你可见到一种特定染色体的许多 UniGene 条目，或使用文库浏览器来观看在 EST 计划中曾经使用过的 cDNA 文库列表。

16.3 在线人类孟德尔遗传(OMIM)[5]

本资源是一个有关已知引起人类疾病的基因的信息数据库，内容十分广泛且经常更新，只需使用一个简单的基于网页的接口和键盘便可以对数据库进行搜索。每次搜索都返回一个所发现命中的总表，点击任何一项都会将你带入实际的条目。每个条目又被分成许多部分，点击文本即可快速浏览，这些部分可能会包括病症的描述、研究结果总结、临床症状、遗传学信息、动物模型、遗传性等。每个条目与许多数据库进行了链接，可以进行广泛地参考，这些链接位于每个条目的顶部。将鼠标的光标放到数据库按钮上，显出链接到那个特定数据库的链接数目，每个列出的参考文献与 PubMed 进行了链接，可以方便地进行交叉参考。OMIM 的每个条目也包含一部分表明是谁创建了本条目、什么时候创建的、一张供稿者列表、一张日期和更新记录者的列表。

16.4 分子模型资源

NCBI 的蛋白质结构数据库称为分子模建数据库(MMDB)，它汇集从 Brookhaven 蛋白质数据库(PDB)[6]得到的三维结构，并将其转换成 ASN.1 格式的记录。MMDB 设计成能够获得生物分子常规的结构信息并对其进一步叙述。使用 Entrez 系统，NCBI 使对结构生物学感兴趣的用户很容易进入本信息。要进入 Entrez 系统，只需在 NCBI 的主页(`http://www.ncbi.nlm.nih.gov/`)上点击 Entrez 并选择三维结构搜索。从这里，你可以进行所有已知三维结构的注释搜索，也可以从 NCBI 的结构主页(`http://www.ncbi.nlm.nih.gov/Structure/`)上进行 Entrez 搜索。进行 Entrez 搜索后，出现一张命中列表，当点击结构概况时，出现与 GenBank 条目、PubMed 和分类学(taxonomy)数据库的超级链接。也可以浏览用 BLAST 计算出的核苷酸邻居(neighbor)和用 Vast[7, 8]算法程序(`http://www.ncbi.nlm.nih.gov/Structure/iucrabs.html`)计算出的结构邻居。使用 NCBI 支持的三种程序中的一种，可以观看任何已解结构的三维图，这些程序是 Rasmol、Mage 和 NCBI 自己的 Cn3D[9]。可以从

http://www.ncbi.nlm.nih.gov/structure/cn3ddown.html 下载 Cn3D，从 Massachusetts 大学的 Rasmol 主页 http://klaatu.oit.umass.edu/microbio/rasmol/index2.htm 下载 Rasmol。

16.5 对分子生物学者有用的工具

除了数据库搜索和序列检索工具外，NCBI 还制作了许多其他工具，供分子生物学者使用。

16.5.1 E-PCR

电子 PCR 或 e-PCR 可使你确定在所查询的序列内是否存在序列标记的位点(sequence tagged sites，STSs)。基于 PCR 的 STSs 是短的 DNA 序列，定位于许多生物的基因组中。STSs 适宜于快速定位新的目的 DNA 片段。目前，这种分析方法最适宜于人类基因组，人类基因组有 45 000 多个 STSs。果蝇有 3203 个、*Bos taurus*(牛)1015 个、*Gallus gallus*(鸡)552 个、*Mus musculus*(家鼠)343 个、*Plasmodium falciparum*(疟疾寄生虫)339 个。许多其他生物都有 STSs，但对每一种的所知甚少。

16.5.2 可读框(ORF)发现程序

对测序资料进行分析时，可读框(ORF)发现程序可能很有用。这种工具能够使你确定在所查询的序列内潜在的可读框所处的位置。将一个序列提交后，在所有的 6 个读框中，所有的可读框以图的形式显示。每个可读框的大小标在图的左侧，先是最大的可读框，最后是最小的可读框。用户可以改变几个参数，例如，所显示的可读框的最小尺寸，以及显示可读框还是每个读框内的终止密码子。点击任何可读框能够进行一次标准的或高级的 BLAST 搜索。

16.5.3 人-鼠同源性图谱

本资源可以使你观看鼠和人基因组之间的同线性区域，点击染色体号可以看到每个染色体，而且用户能够看到人和鼠基因的染色体定位。每条染色体的资料显示在一张表格内，列出基因的名称和相关的定位资料。与每个基因相关的信息包括：

(1) Genethon 图定位。

(2) 对基因进行定位的方法。

(3) 在 OMIM(原位栏)中叙述的疾病基因和其他表达基因的细胞遗传学定位。

(4) 与其他基因组信息学站点的超级链接，鼠基因为 Jackson 实验室(点击鼠基因)，人类基因为 OMIM(点击人类基因)。

(5) 相应的鼠或人染色体。

(6) 放射杂交定位区域，能将用户连接到人类基因组的基因图(gene map)上。随后超级链接到标志物上，将提供标志物的详细情况，并可以超级链接到实际的人类基因图区域(通过选择 Genethon 图基因座间隔定义的间隔)。通过选择一个间隔，能够呈现此区域内的所有 Marker。

(7) 鼠或人同染色体基因的定位。

(8) 杂交栏，表明哪个实验室对一个已知的杂交进行了基因定位。

16.5.4 分类学

本数据库能够使你看到在 NCBI 遗传学数据库中有至少一种核苷酸或蛋白质序列的所有生物的名称。选择分类浏览器，可以对数据库进行搜索。进入任何一部分分类学叙述，或甚至于普通的名称，都可以分级地看到这种分类。例如，键入“Mus”，将列出在 GenBank 中有蛋白质或基因的所有亚种。点击任何亚种，将为用户提供有关此亚种的进一步的信息，并将列出 DNA 或蛋白质序列的号或三维结构。使用 Entrez 的注释搜索可以在 GenBank 中找到，使用 Entrez 系统可以观看每组的序列。

16.6 小结

随着计算技术的进步，以及收集到新的有价值的生物学信息，数据库的数量和用于分析其所含资料的工具也将改变。从本文开始写作时到将小结归纳到一起(约 6 个月)，GenBank 又出了 4 个版本、出了两种新的 BLAST 搜索引擎——PHI-BLAST 和生物特异性 BLAST、一个人类遗传变异数据库(dbSNP)和一种新版本的 Cn3D。在不远的将来，NCBI 将为研究团体提供大量的信息和计算工具，为分子生物学者提供许多其他的工具。

(李慎涛　译)

参 考 文 献

[1] Benson, D. A., Boguski, M. S., Lipman, D. J., Ostell, J., and Ouellette, B. F. (1998) GenBank. *Nucleic Acids Res.* **26,** 1-7.

[2] Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990) Basic local alignment search tool. *J. Mol. Biol.* **215,** 403-410.

[3] Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.***25,** 3389-3402.

[4] Tatusov, R. L., Koonin, E. V., and Lipman, D. J. (1997) A genomic perspective on protein families. *Science* 24, 631-637.

[5] Online Mendelian Inheritance in Man, OMIM (TM). (1997) Center for Medical Genetics, Johns Hopkins University

(Baltimore, MD) and National Center for Biotechnology Information, National Library of Medicine (Bethesda, MD).

[6] Abola, E. E., Bernstein, F. C., Bryant, S. H., Koetzle, T. F., and Weng, J. C. (1987) Protein data bank, in *Crystallographic databases: information content, software systems, scientific applications.* (Allen, F. H., Bergerhoff, G, Sievers R., eds.) International Union of Crystallography, Chester, Cambridge, UK, pp. 107-132.

[7] Madej, T., Gibrat, J-F., and Bryant, S. H. (1995) Threading a database of protein cores. *Protein Struct. Funct. Genet.* **23,** 356-369.

[8] Gibrat, J-F., Madej, T., and Bryant, S. H. (1996) Surprising similarities in structure comparison. *Curr. Opin. Struct. Biol.* **6,** 377-385.

[9] Hogue, C. W. V. (1997) Cn3D: a new generation of three-dimensional molecular structure viewer. *Trends Biochem. Sci.* **22,** 314-316.

17 EBI 上的资源

Patricia Rodriguez-Tomé

17.1 引言

欧洲生物信息学研究所(European Bioinformatics Institute，EBI)是 EMBL 的分部，位于英国 Hinxton 的 Wellcome Trust Genome Campus。EBI 维护和发布 EMBL 核酸序列数据库、欧洲原始核酸序列数据资源库、SwissProt 蛋白质序列数据库[与瑞士生物信息学协会(Swiss Institute for Bioinformatics，SIB)的 Amos Bairroch 合作]、TrEMBL(SwissProt 的附属数据库，由 EMBL 数据库编码序列翻译而来的蛋白质序列数据库)、分子结构数据库(Molecular Structure Database，MSD)[与 Brookhaven 国家实验室(Brookhaven，纽约)的蛋白质三维结构数据库(Protein Data Bank ，PDB)合作]、放射杂交数据库(Radiation Hybrid database，RHdb)和由其他组织合作产生的分子生物学数据库。EBI 还提供网络服务，通过因特网、其万维网界面和 ftp 服务器可以访问最新收集到的数据，同时也提供数据库和序列相似性的搜索工具。

17.2 数据库

17.2.1 EMBL——核酸序列数据库

EMBL 核酸序列数据库[1]是 EBI 的主要领域，初创于 20 世纪 80 年代，它是收集、组织和发布核酸序列数据和相关信息的数据库，自 1982 年以来，该项工作由 EBI 与 GenBank[2](NCBI)以及随后加入的 DDBJ(DNA Database of Japan)[3]共同完成，每个中心收集和更新数据，并每天进行交换。

由于许多研究机构使用了高通量的测序技术，所以目前每分钟就可以向 EMBL 提交一个序列，EMBL 正在和产生大量数据的主要测序项目进行协作(图 17.1)。

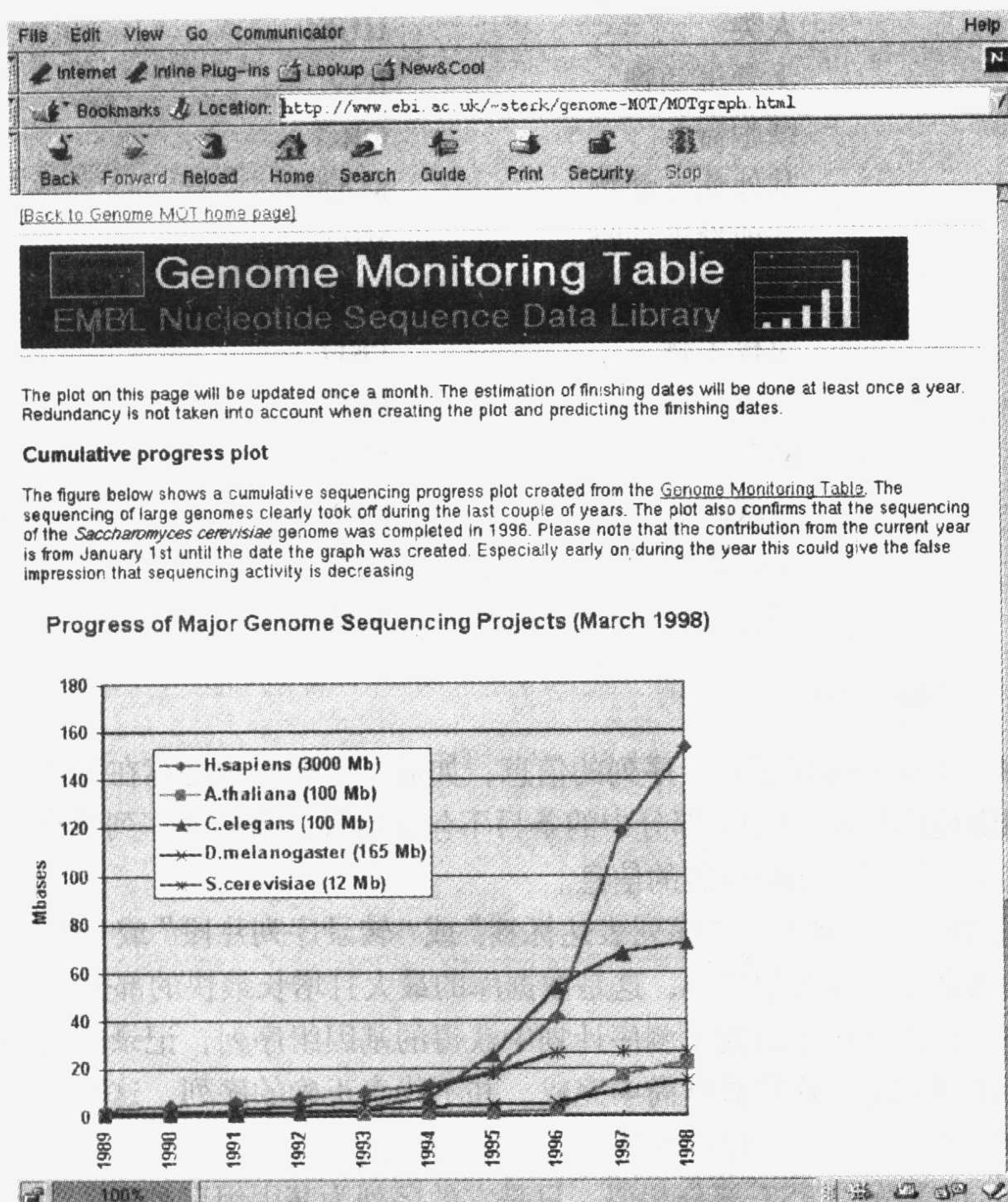


图 17.1　基因组监视表，报告基因组中心的测序进展

17.2.1.1　数据的分类

除了像 HTG(高通量基因组序列)或 GSS(基因组调查序列)这类极少数的情况以外，EMBL 数据库主要根据分类学来分类，使用如下的 3 个字母的代码来定义分类：

分类	代码
噬菌体	PHG
构建的序列	CON
表达序列标签	EST
真菌	FUN
高通量基因组	HTG
基因组调查序列	GSS

人类	HUM
无脊椎动物	INV
细胞器	ORG
其他哺乳动物	MAM
其他脊椎动物	VRT
植物	PLN
原核生物	PRO
啮齿动物	ROD
STS	STS
合成的	SYN
未分类的	UNC
病毒	VRL

17.2.1.2 特殊部分

CON：该部分保存了很长序列的信息，如通过数据库中已存在序列组装的全基因组和染色体片段。CON 部分中的条目不包括特性、表格或序列数据，但是它们包括该全长序列是如何拼装的信息。

EST：该部分包括称为“序列表达标签”或“转录序列片段”或“部分 cDNA”的序列，独立于其分类学分类，这是数据库的最大且增长最快的部分。

HTG：该部分用于高通量测序计划中获得的基因组序列，记录由长的序列组成。需要注意的是，这些数据尚未完成，并不代表正确的序列，这部分数据的发布基于如下考虑：随着工作的继续，这些序列会发生改变。

GSS：该部分与 EST 部分相似，只是它的序列为基因组序列而不是 cDNA。条目包括从随机基因组调查序列、外显子捕捉结果序列、Alu PCR 序列、BAC 或 YAC 末端克隆序列由“单向通读”(“single pass reads”)所产生的序列数据。

Patents：该部分在上表中没有列出，是来源于专利局的序列。欧洲专利局(European Patent Office)在专利申请日的 18 个月后向外公布数据，EMBL 数据库会立即收录这些数据并对外公布。

NEW：该部分不是数据库的一个特殊部分，只包括自上一个完全版本以来所收集到的所有数据。该部分每天更新，在版本冻结时限(即当创建一个版本时)会消失，但是当有新数据提交后，又会重新创建。

17.2.1.3 数据库条目结构

数据库条目用 EMBL 平面文件格式(图 17.2)发布，该格式为大多数序列分析软件包所支持，该格式提供了一种方便于读者的结构，由不同的行类型(line type)组成，用来记录组成一个条目的不同类型的数据。典型的数据库条目包含下述行类型：

```
ID   HSIGHAF    standard; RNA; HUM; 1089 BP.
XX
AC   J00231;
XX
NI   g185041
XX
DT   13-JUN-1985 (Rel. 06, Created)
DT   17-DEC-1994 (Rel. 42, Last updated, Version 6)
XX
DE   Human Ig gamma3 heavy chain disease OMM protein mRNA.
XX
KW   C-region; gamma heavy chain disease protein;
KW   gamma3 heavy chain disease protein; heavy chain disease; hinge exon;
KW   immunoglobulin gamma-chain; immunoglobulin heavy chain;
KW   secreted immunoglobulin; V-region.
XX
OS   Homo sapiens (human)
OC   Eukaryota; Metazoa; Chordata; Vertebrata; Mammalia; Eutheria; Primates;
OC   Catarrhini; Hominidae; Homo.
XX
RN   [1]
RP   1-1089
RX   MEDLINE; 82247835.
RA   Alexander A., Steinmetz M., Barritault D., Frangione B., Franklin E.C.,
RA   Hood L., Buxbaum J.N.;
RT   "gamma Heavy chain disease in man: cDNA sequence supports partial gene
RT   deletion model";
RL   Proc. Natl. Acad. Sci. U.S.A. 79:3260-3264(1982).
XX
DR   GDB; 119339; IGHG3.
DR   GDB; G00-119-339.
DR   IMGT/LIGM; J00231; Release 98.03.
DR   SWISS-PROT; P01860; GC3_HUMAN.
XX
CC   The protein isolated from patient OMM is a gamma heavy chain
CC   disease (HCD) protein. It has a large 5' internal deletion
CC   consisting of most of the variable region and the entire ch1
CC   domain. [1] suggests that the protein abnormality is from a partial
CC   gene deletion rather than from defective splicing. NCBI gi: 185041
XX
FH   Key             Location/Qualifiers
FH
FT   source          1. .1089
FT                   /organism="Homo sapiens"
FT   mRNA            <1. .1089
FT                   /note="gamma3 mRNA"
FT   CDS             23. .964
FT                   /codon_start=1
FT                   /db_xref="PID:g567112"
FT                   /db_xref="SWISS-PROT:P01860"
FT                   /note="OMM protein (Ig gamma3) heavy chain; NCBI gi:
FT                   567112"
FT                   /gene="IGHG3"
FT                   /map="14q32.33"
FT                   /translation="MKXLWFFLLLVAAPRWVLSQVHLQESGPGLGKPPELKTPLGDTTH
FT                   TCPRCPEPKSCDTPPPCPRCPEPKSCDTPPPCPRCPEPKSCDTPPPCPXCPAPELLGGP
FT                   SVFLFPPKPKDTLMISRTPEVTCVVVDVSHEDPXVQFKWYVDGVEVHNAKTKLREEQYN
FT                   STFRVVSVLTVLHQDWLNGKEYKCKVSNKALPAPIEKTISKAKGQPXXXXXXXXXXXXXE
FT                   EMTKNQVSLTCLVKGFYPSDIAVEWESNGQPENNYNTTPPMLDSDGSFFLYSKLTVDKS
FT                   RWQQGNIFSCSVMHEALHNRYTQKSLSLSPGK"
FT   sig_peptide     26. .79
FT                   /codon_start=1
FT                   /note="OMM protein signal peptide"
FT   mat_peptide     80. .961
FT                   /codon_start=1
FT                   /note="OMM protein mature peptide"
XX
SQ   Sequence 1089 BP; 240 A; 358 C; 271 G; 176 T; 44 other;
     CCTGGACCTC CTGTGCAAGA ACATGAAACA NCTGTGGTTC TTCCTTCTCC TGGTGGCAGC        60
     TCCCAGATGG GTCCTGTCCC AGGTGCACCT GCAGGAGTCG GGCCCAGGAC TGGGGAAGCC       120
     TCCAGAGCTC AAAACCCCAC TTGGTGACAC AACTCACACA TGCCCACGGT GCCCAGAGCC       180
     CAAATCTTGT GACACACCTC CCCCGTGCCC ACGGTGCCCA GAGCCCAAAT CTTGTGACAC       240
     ACCTCCCCCA TGCCCACGGT GCCCAGAGCC CAAATCTTGT GACACACCTC CCCCGTGCCC       300
     NNNGTGCCCA GCACCTGAAC TCTTGGGAGG ACCGTCAGTC TTCCTCTTCC CCCCAAAACC       360
     CAAGGATACC CTTATGATTT CCCGGACCCC TGAGGTCACG TGCGTGGTGG TGGACGTGAG       420
     CCACGAAGAC CCNNNNGTCC AGTTCAAGTG GTACGTGGAC GGCGTGGAGG TGCATAATGC       480
     CAAGACAAAG CTGCGGGAGG AGCAGTACAA CAGCACGTTC CGTGTGGTCA GCGTCCTCAC       540
     CGTCCTGCAC CAGGACTGGC TGAACGGCAA GGAGTACAAG TGCAAGGTCT CCAACAAAGC       600
     CCTCCCAGCC CCCATCGAGA AAACCATCTC CAAAGCCAAA GGACAGCCCN NNNNNNNNNN       660
     NNNNNNNNNN NNNNNNNNNN NNNNNGAGGA GATGACCAAG AACCAAGTCA GCCTGACCTG       720
     CCTGGTCAAA GGCTTCTACC CCAGCGACAT CGCCGTGGAG TGGGAGAGCA ATGGGCAGCC       780
     GGAGAACAAC TACAACACCA CGCCTCCCAT GCTGGACTCC GACGGCTCCT TCTTCCTCTA       840
     CAGCAAGCTC ACCGTGGACA AGAGCAGGTG GCAGCAGGGG AACATCTTCT CATGCTCCGT       900
     GATGCATGAG GCTCTGCACA ACCGCTACAC GCAGAAGAGC CTCTCCCTGT CTCCGGGTAA       960
     ATGAGTGCCA TGGCCGGCAA GCCCCCGCTC CCCGGGCTCT CGGGGTCGCG CGAGGATGCT      1020
     TGGCACGTAC CCCGTGTACA TACTTCCCAG GCACCCAGCA TGGAAATAAA GCACCCAGCG      1080
     CTGCCCTGG                                                              1089
//
```

图 17.2 一个 EMBL 条目

AC：包含那个条目的一个独一无二的识别号(identifier)(登录号，accession number)。

DE：一个简洁的描述行。

OS：来源生物的分类学描述。

参考文献信息：RA 行为作者信息、RT 行为文章标题、RL 行为期刊信息。

NI：为核酸序列的识别号，NI 在序列本身有所变化时也随之改变，但是登录号保持不变。该识别号将被 1999 年的一个适当的版本号替代。

FT：特性表，用来描述编码区和其他具有生物学意义的位点的位置，特性表的下面是统一的 DDBJ / EMBL / GenBank 特性表定义，以一种可以在 URL http://www.ebi.ac.uk/ebi_docs/embl_db/ft/feature_table.html 上获得的文件形式描述。

SQ：序列本身。

1) 登录号

数据库中每个条目都有一个独一无二的识别号，即它的登录号。以前的条目中该识别号(在 AC 行内)由一个前缀字母加 5 个数字(1+5 格式)组成，新的条目中则由 2 个前缀字母加 6 个数字(2+6 格式)组成。由于基因组计划正在产生大量的数据，登录号的空间必须扩展。在 EMBL、DDBJ 和 GenBank 数据库中该登录号是不变的，在三个数据库中都会指向同一个条目。

2) 蛋白质识别号(PID)

蛋白质识别号(protein identifier)指某一特定条目的编码序列的翻译产物。该 PID 在该条目的 CDS 特性中有所描述。只要该条目编码序列的翻译不变，该识别号就保持不变，这样，它可以被外部数据库用作识别号，根据此识别号可以建立交叉引用。

17.2.1.4 数据提交(data submission)

EBI 提供了多种提交新序列数据的方法：

(1) WEBin(图 17.3)为网上序列提交工具，可以在 URL http://www.ebi.ac.uk/submission/webin.html 上获得，该工具带领提交者通过所有的步骤，以交互式的、简便的方式来提交序列数据和描述性信息。

(2) Sequin 是由 NCBI 开发的向 EMBLI/GenBank /DDBJ 数据库提交条目的工具，该软件可以在 EBI 的匿名 ftp 服务器 ftp://ebi.ac.uk/pub/software/sequin/上获得。

(3) 对于只能够通过 e-mail 进入网络的提交者，还可以得到一张数据提交表格，该表格索要创建一个数据库条目所需的一切信息，用电子邮件通过 EBI 的服务器可以得到该表格，用户可以向 netserv@ebi.ac.uk 发送以下内容的邮件：

```
GET DOC:DATASUB.TXT
```

图 17.3　WEBin，EMBL 的网上提交工具

(4) 如果用户要提交超过 25 条以上的条目，使用交互式工具将非常繁琐，EBI 鼓励提交者在提交数据前和数据库联系，数据库的工作人员将帮助使数据提交尽可能的方便。

当序列已被提交给 EMBL 且被接受，数据库的工作人员将给提交者提供一个独一无二的识别号，即识别该条目的一个登录号。该条目将以同一个登录号自动出现在 GenBank 和 DDBJ，因为向 3 大数据库的任何一个提交，它们每天都互相转寄。

17.2.1.5　数据的秘密性

提交到数据库的序列可以在处理后马上发布，或在发表时公布，这取决于提交者的意愿。对于不立即公开的条目则不转寄到其他的数据库，直至提交者要求

公布或出版(不管哪种先到)。

17.2.1.6 数据更新

保持条目正确和最新的唯一的方式是作者将其新的发现或改正通知数据库的工作人员，信息可以通过以下方式之一进行交流：

(1) 网络：地址为：http://www.ebi.ac.uk/ebi_docs/update.html。

(2) e-mail: update@ebi.ac.uk。

(3) FTP：在 ftp://ebi.ac.uk/pub/databases/embl/release/update.doc 获得一个文件，也可以将更新传真至数据库。

欢迎使用者指出在数据库中发现的错误，但是应当知道，只有原始提交者有权对序列进行更新和对主要注解进行修改，数据更新在 3 个数据库间同步进行。

17.2.1.7 数据库发行

EMBL 数据库每季度以 CD-ROM 的形式进行公布和发行，也可以在 FTP 服务器(ftp://ftp.ebi.ac.uk/pub/databases/embl)上获得。数据库中新的和更新的条目每天都向服务器中添加，这就可以让使用者及时在网络上得到这些数据，在 EBI 查询 EMBL 核酸序列数据库的特殊服务将在本章的第二部分叙述。

17.2.2 SwissPort 数据库

SwissPort[4]是一个注释蛋白质序列的数据库，是由 EBI 和日内瓦大学(瑞士)医学生物化学系合作创建的。它包含高质量的注解数据，是非冗余的，并与许多其他数据库交互引用。为了标准化的目的，SwissPort 的格式尽可能和 EMBL 条目的格式相一致。一个 SwissPort 的实例见图 17.4。

一个典型的条目包括序列数据、参考文献、分类学信息和注解本身，包括如下条目：

- 蛋白质的功能。
- 翻译后修饰。
- 结构域和位点。
- 二级结构。
- 四级结构。
- 与其他蛋白质的相似性。
- 和该蛋白质缺陷相关的疾病。
- 序列的矛盾(conflict)和变异体(variant)等。

```
ID   GC3_HUMAN      STANDARD;      PRT;   290 AA.
AC   P01860;
DT   21-JUL-1986 (REL. 01, CREATED)
DT   21-JUL-1986 (REL. 01, LAST SEQUENCE UPDATE)
DT   01-FEB-1991 (REL. 17, LAST ANNOTATION UPDATE)
DE   IG GAMMA-3 CHAIN C REGION (HEAVY CHAIN DISEASE PROTEIN) (HDC).
GN   IGHG3.
OS   HOMO SAPIENS (HUMAN).
OC   EUKARYOTA; METAZOA; CHORDATA; VERTEBRATA; TETRAPODA; MAMMALIA;
OC   EUTHERIA; PRIMATES.
RN   [1]
RP   SEQUENCE (DISEASE PROTEIN WIS).
RX   MEDLINE; 81021548.
RA   FRANGIONE B., ROSENWASSER E., PRELLI F., FRANKLIN E.C.;
RL   BIOCHEMISTRY 19:4304-4308(1980).
RN   [2]
RP   NORMAL GAMMA-3 CHAINS, REVISIONS TO 12-97 OF PROTEIN WIS.
RX   MEDLINE; 77118561.
RA   MICHAELSEN T.E., FRANGIONE B., FRANKLIN E.C.;
RL   J. BIOL. CHEM. 252:883-889(1977).
RN   [3]
RP   DISEASE PROTEIN ZUC, REVISIONS TO 59-289 OF PROTEIN WIS.
RX   MEDLINE; 77021516.
RA   WOLFENSTEIN-TODEL C., FRANGIONE B., PRELLI F., FRANKLIN E.C.;
RL   BIOCHEM BIOPHYS. RES. COMMUN. 71:907-914(1976).
RN   [4]
RP   SEQUENCE FROM N.A. (DISEASE PROTEIN OMM).
RX   MEDLINE; 82247835.
RA   ALEXANDER A., STEINMETZ M., BARRITAULT D., FRANGIONE B.,
RA   FRANKLIN E.C., HOOD L., BUXBAUM J.N.;
RL   PROC. NATL. ACAD. SCI. U.S.A. 79:3260-3264(1982).
CC   -!- SUBUNIT: DIMER LINKED BY 12 DISULFIDE BONDS; IT HAS AN EXTRA
CC       INTERCHAIN DISULFIDE BOND AT POSITION 7 IN ADDITION TO THE 11
CC       NORMALLY PRESENT IN THE HINGE REGION.
CC   -!- THE HEAVY CHAIN DISEASE PROTEIN WIS IS SHOWN.
CC   -!- THE SEQUENCE OF RESIDUES 42-76 WAS TAKEN FROM THE REF. 2.
CC   -!- DISEASE PROTEIN WIS IS LACKING MOST OF THE V REGION AND ALL OF THE
CC       CH1 REGION.
CC   -!- DISEASE PROTEIN ZUC LACK MOST OF THE V REGION, ALL OF THE CH1
CC       REGION, AND PART OF THE HINGE COMPARED WITH NORMAL GAMMA-3 HEAVY
CC       CHAINS.
CC   -!- DISEASE PROTEIN OMM MAY REPRESENT AN ALLELIC FORM OR ANOTHER GAMMA
CC       CHAIN SUBCLASS.
CC   -!- THE HINGE REGION IN GAMMA-3 CHAINS IS ABOUT FOUR TIMES AS LONG
CC       AS IN OTHER GAMMA CHAINS AND CONTAINS THREE IDENTICAL 15-RESIDUE
CC       SEGMENTS PRECEDED BY A SIMILAR 17-RESIDUE SEGMENT (12-28).
DR   EMBL; J00231; G567112; ALT_SEQ.
DR   PIR; A02149; G3HUWI.
DR   HSSP; P01857; 1FC1.
DR   MIM; 147120; -.
DR   PROSITE; PS00290; IG_MHC; 1.
KW   IMMUNOGLOBULIN C REGION; GLYCOPROTEIN.
FT   DOMAIN       12     73       HINGE.
FT   DOMAIN       74    183       CH2.
FT   DOMAIN      184    289       CH3.
FT   REPEAT       29     43
FT   REPEAT       44     58
FT   REPEAT       59     73
FT   MOD_RES       1      1       PYRROLIDONE CARBOXYLIC ACID.
FT   CARBOHYD      6      6
FT   DISULFID      7      7       INTERCHAIN (WITH HEAVY CHAIN DIMER).
FT   DISULFID     24     24       INTERCHAIN (WITH HEAVY CHAIN DIMER).
FT   DISULFID     27     27       INTERCHAIN (WITH HEAVY CHAIN DIMER).
FT   DISULFID     33     33       INTERCHAIN (WITH HEAVY CHAIN DIMER).
FT   DISULFID     39     39       INTERCHAIN (WITH HEAVY CHAIN DIMER).
FT   DISULFID     42     42       INTERCHAIN (WITH HEAVY CHAIN DIMER).
FT   DISULFID     48     48       INTERCHAIN (WITH HEAVY CHAIN DIMER).
FT   DISULFID     54     54       INTERCHAIN (WITH HEAVY CHAIN DIMER).
FT   DISULFID     57     57       INTERCHAIN (WITH HEAVY CHAIN DIMER).
FT   DISULFID     63     63       INTERCHAIN (WITH HEAVY CHAIN DIMER).
FT   DISULFID     69     69       INTERCHAIN (WITH HEAVY CHAIN DIMER).
FT   DISULFID     72     72       INTERCHAIN (WITH HEAVY CHAIN DIMER).
FT   CARBOHYD    140    140
FT   MOD_RES     290    290       REMOVED POST-TRANSLATIONALLY.
FT   VARIANT     126    127       QV -> EB (IN ZUC).
FT   VARIANT     134    134       P -> L (IN OMM).
FT   VARIANT     139    139       F -> Y (IN OMM).
FT   VARIANT     182    182       T -> A (IN OMM).
FT   VARIANT     227    227       S -> N (IN OMM).
FT   VARIANT     227    227       MISSING (IN ZUC).
FT   VARIANT     279    279       F -> Y (IN OMM).
SQ   SEQUENCE   290 AA;  32331 MW;  C5E7BE05 CRC32;
     QMQGVNCTVS SELKTPLGDT THTCPRCPEP KSCDTPPPCP RCPEPKSCDT PPPCPRCPEP
     KSCDTPPPCP RCPAPELLGG PSVFLFPPKP KDTLMISRTP EVTCVVVDVS HEDPEVQFKW
     YVDGVQVHNA KTKPREQQFN STFRVVSVLT VLHQNWLDGK EYKCKVSNKA LPAPIEKTIS
     KTKGQPREPQ VYTLPPSREE MTKNQVSLTC LVKGFYPSDI AVEWESSGQP ENNYNTTPPM
     LDSDGSFFLY SKLTVDKSRW QQGNIFSCSV MHEALHNRFT QKSLSLSPGK
//
```

图 17.4　一个 SwissProt 条目

SwissProt 条目是由 EMBL 上的序列翻译、从文献中摘录或者由研究者直接提交而来。为了创建注解，SwissProt 的维护者不仅查阅作者引用的参考文献，而且还查阅相关的文献，以定期更新该蛋白质家族或该蛋白质类型的注解。外部的合作者为特定的结构域提供意见。在 SwissProt，注解主要在注释行(CC 行)、特性表(FT 行)以及关键字行(KW 行)中出现。

通过合并不同参考文献报道的同一蛋白质的不同序列，实现了最小冗余，对于不同文献报道间的矛盾，在条目的特性表中有所指示。

17.2.2.1　和其他数据库的整合

目前 SwissProt 和 30 个不同的数据库交叉引用，交叉引用出现在 DR 部分，其格式如下：

```
Database_name;primary_identifier;secondary_identifier
```

一个 SwissProt 条目可以指向同一数据库中的多个条目。

17.2.2.2　数据的提交

SwissProt 用户要提交新的序列数据或更新条目，应当和 EBI 的 e-mail 地址联系：datasubs@ebi.ac.uk。

17.2.2.3　TrEMBL

目前，由于核酸序列数据库的增长太快，以至于在保持同样的质量标准下 SwissProt 的手工注解处理非常困难。在 1999 年的 3 月，EMBL 有 320 万个条目，且该数字每年都要增加一倍。SwissProt 则包含 77 977 个条目。

在1997 年的早期，引入了 TrEMBL(EMBL 的核酸序列数据库的翻译)，TrEMBL 作为一个计算机注解条目的数据库，是按照 SwissProt 的格式，从 EMBL 的所有编码序列(CDS)翻译中衍生出来的，但是不包括已经在 SwissProt 条目中存在的 CDS。TrEMBL 中的条目不断地与 SwissProt 条目合并。

指向一个 EMBL 条目的 SwissProt 和 TrEMBL 条目中的 DR 行引用 EMBL 的登录号作为原始的识别号，PID 作为二次识别号。

17.2.2.4　数据的发布

每个季度，SwissProt 和 TrEMBL 都用光盘发布，每个数据库的完整版本都可以从 EBI 的匿名 ftp 服务器得到，其地址为：ftp://ftp.ebi.ac.uk/pub/databases/swissprot 和 ftp://ftp.ebi.ac.uk/pub/databases/trembl。

17.2.3　MSD 数据库

分子结构数据库(Molecular Structure Database，MSD)是围绕蛋白质三维结构

数据库(PDB，Brookhaven，NY)发展而来。PDB[5]是一个用实验方法确定的生物大分子三维结构存档文件，服务于全球的研究机构、教育者和学生，该存档文件包含了原子坐标、文献引用、一级和二级结构信息、晶体结构和核磁共振(NMR)实验数据。该数据库可以在 `http://www2.ebi.ac.uk/pdb/`得到，这是北美万维网站点的一个镜像。

17.2.3.1 数据提交

现在用户可以向 EBI 或者 Brookhaven 的 PDB 提交数据。欧洲的使用者在 EBI 的欧洲网站上可以非常便利地提交数据。EBI 和 Brookhaven 实验室的合作也包括开发新的提交软件工具 AutoDep(图 17.5)，可以在 `http://autodep.ebi.ac.uk/autodep-basepage.shtml` 得到。

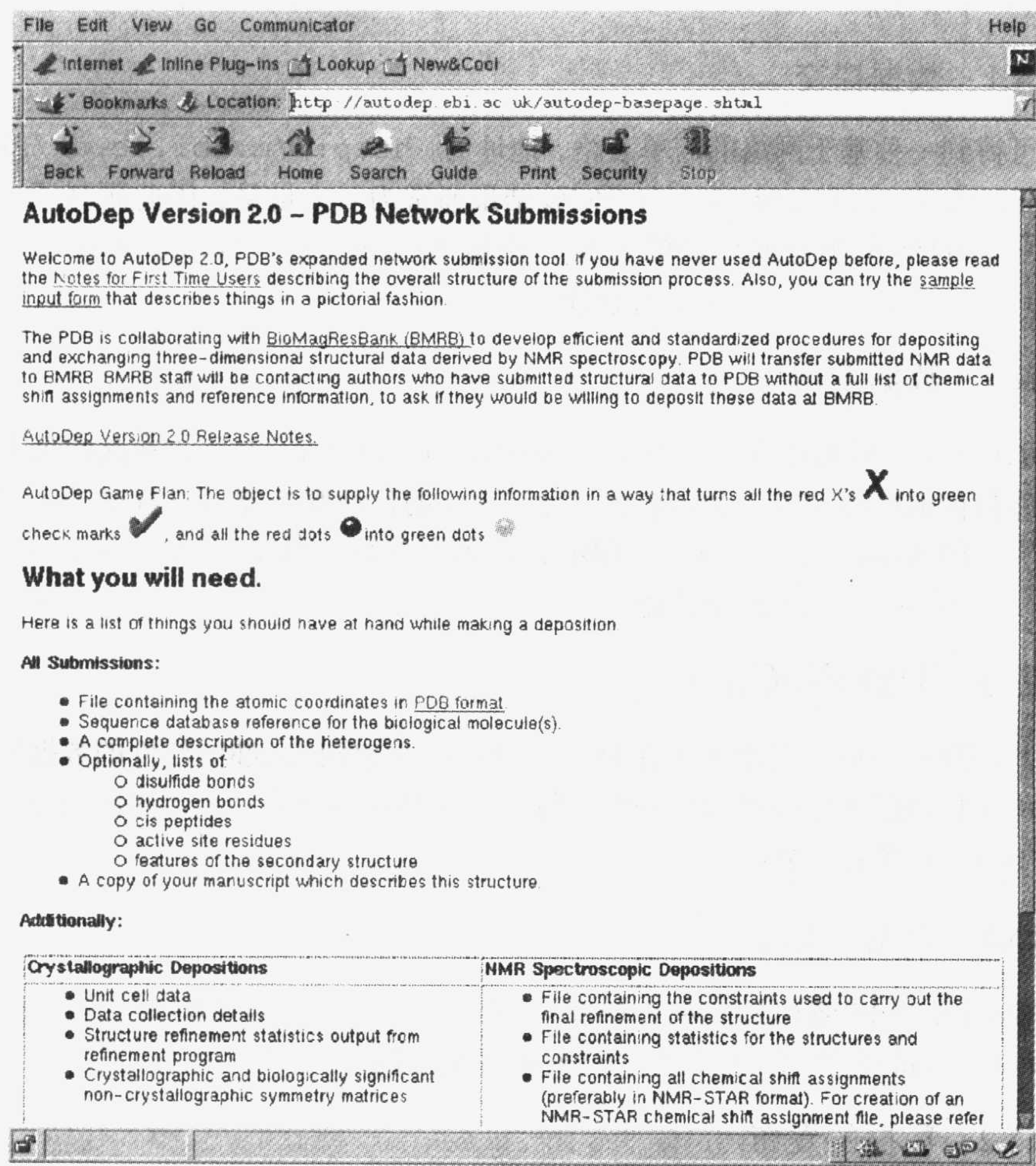


图 17.5 AutoDep——PDB/MSD 的网上提交工具

17.2.3.2 数据的发布

PDB 在每个季度发布完整的数据库版本，也可以在 EBI 的匿名服务器 ftp://ftp.ebi.ac.uk/pub/databases/pdb 上获得，每周更新一次的数据同样可以在上述的服务器上获得。

17.2.4 放射杂交数据库

放射杂交数据库(Radiation Hybrid Database，RHdb)[6]是一个和放射杂交图谱相关的原始数据的数据库。RHdb 储存了有关 panels、实验条件、STSs，以及分析的实验结果、图谱信息和图谱的资料。RHdb 的一个重要方面是和其他数据库广泛的交叉参考，RHdb 中的每个条目都有一个独一无二的识别号，其形式为 RH*n*，*n* 代表一个数字。

17.2.4.1 数据提交

可得到一种基于网络的简单表格，地址为：http://m.ebi.ac.uk/RHdb/submission_form.html。由于实验结果通常为一大批量，因此研制了一种带标签区域的格式来处理大量的数据，该格式在 http：//www.ebi.ac.uk/RHdb/data_input.html 上有解释。

17.2.4.2 数据发布

RHdb 的主页地址为：http：//www.ebi.ac.uk/RHdb，在此网页上有该数据库目前的所有信息，通过各种基于 Java 的代理服务器可以直接进入该数据库。完整的版本(和更新)可以在 EBI 的匿名服务器 ftp://ftp.ebi.ac.uk/pub/databases/RHdb 上获得。

17.2.5 The BioCatalog

The BioCatalog[7]是个关于在分子生物学和遗传学感兴趣的软件的数据库。不同的程序按照感兴趣的结构域分组，每个条目都有一个独一无二的识别号，其形式为 BC*n*，*n* 代表一个数字。

17.2.5.1 数据的提交

使用基于网络(地址为：http://www.ebi.ac.uk/biocat/biocat_form.html)的表格可以向 BioCatalog 添加新的信息。

17.2.5.2 数据的发布

BioCatalog 以 ASCII 文件的形式免费发布，可以在 EBI 的匿名 ftp 服务器(地

址为：ftp://ftp.ebi.ac.uk/pub/databases/biocat)上获得，该目录可以从 EBI 网络服务器(地址为：http://www.ebi.ac.uk/biocat)上得到，由结构域/子类产生列表。使用者可以按照链接到达原始网站来获得软件，也可以使用关键词在 BioCatalog 上检索。

17.3 EBI 网络服务

除了维护各种数据库以外，EBI 也向外部用户提供了大量的免费网络服务。

17.3.1 匿名的 ftp 服务器

这是获得 EBI 数据库完整版本和更新版本，以及软件库中的软件的主要途径，主要的服务器在 ftp://ftp.ebi.ac.uk/pub/，在此用户能够浏览不同的目录。

- 数据库：所有数据库发布。
- Doc：EBI 文档。
- 帮助：数据库和服务的帮助文件。
- 软件：软件压缩包。

17.3.2 网络文件服务器

文件服务器能够通过 e-mail 进入 EBI 维护的全部数据库、公开的结构域软件和文档。向文件服务器地址(netserv@ebi.ac.uk.)发送的电子邮件中发送一条命令，便可以从服务器得到条目，向服务器发送 help 文字信息的邮件将会得到如何使用服务器的最新帮助文件。

17.3.3 万维网服务器

EBI 的主页(地址为：http://www.ebi.ac.uk/)能够进入所有的 EBI 服务和数据库，该主页随着技术的进步和服务的变更而不断更新，但是其理念是始终不变的，即列出 EBI 不同领域的信息，如基础研究、服务、工业程序和一般的信息，所有该章描述的数据库和服务都可以在服务项目中找到。

17.3.3.1 序列检索和相关信息：SRS 系统

SRS[8]是序列检索系统(Sequence Retrieval System)，该软件包起初为方便查询一个序列条目的注释部分，现在已经发展成为一个完整的工作平台，用于序列分析，并整合了常见的序列分析软件。为了更好地检索，该系统开发时考虑到了不同数据库之间的链接(交叉参考)。

SRS 网站是一个对用户非常友好的 SRS 的网络界面，其地址为：http://srs.ebi.ac.uk/。

该网页上的按钮和不同的文档链接：SRS Manual(提供使用该系统的全部信息)、SRS WorldWide(列出了网络上的SRS服务器)、SRS newsgroup[能够进入SRS特异的新闻组(bionet.software.srs)]、SRS Developers(列出了SRS开发队伍)。

我们将举例说明SRS的基本用途，但是SRS用户手册中有关于SRS所有特性的全部最新信息。

要使用SRS，点击Start按钮，用户便可以开始。一次使用能够在整个当前连接中维持用户的选择和配置。

l) Top Page

接下来的网页叫“SRS top page”，该网页用来选择要搜索的数据库。用户点击数据库名称附近的复选框即可选中数据库，数据库名称的本身与提供有关用于搜索各个数据库的索引信息的网页链接，当需要改变数据库的搜索设置时，则需要返回到“this top page”。在图17.6的示例中，选中了EMBL和EMBLnew(自上一版本以来的累积更新)数据库。

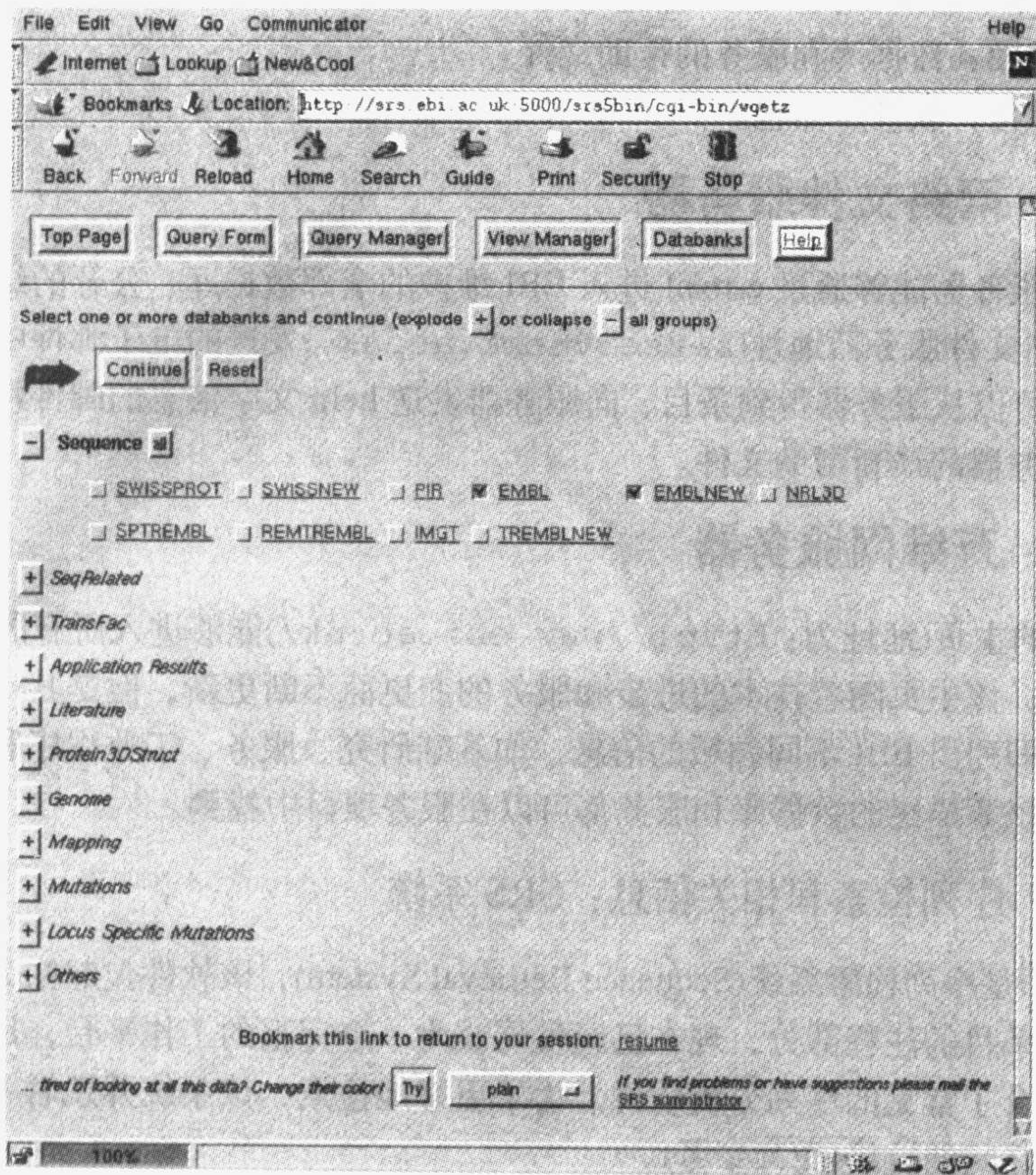


图17.6 SRS主界面

此图中EMBL和EMBL new数据库都被选择用于搜索

点击 reset 按钮将取消此网页上的所有选择，SRS 网页上所有的 reset 按钮都有同样的功能。点击 continue 按钮出现“查询表格页”。

2) 查询表格页

在文本输入字段(图 17.7)左侧的 choice 按钮列出了查询所选中的数据库所有可用的字段(field)，在此例子中，“Description”、“Organism”、“Description”和“All Text”字段可用。该列表取决于 SRS 服务器检索了哪些字段，不同的站点会检索不同的字段，所以对于不同的 SRS 服务器，就是对于同一数据库该列表也会有所不同。如果从顶页选择了“show only fields that selected databases have in common”复选框，那么只列出索引字段的常用子集。如果在两个数据库中并非所有的检索字段都出现，用户可以在搜索字段输入其他相关的关键词，“include fields in output”列表能够使用户选择在“查询结果页”上显示哪个字段。点击“do query”按钮将执行查询，下一个网页是“查询结果页”。

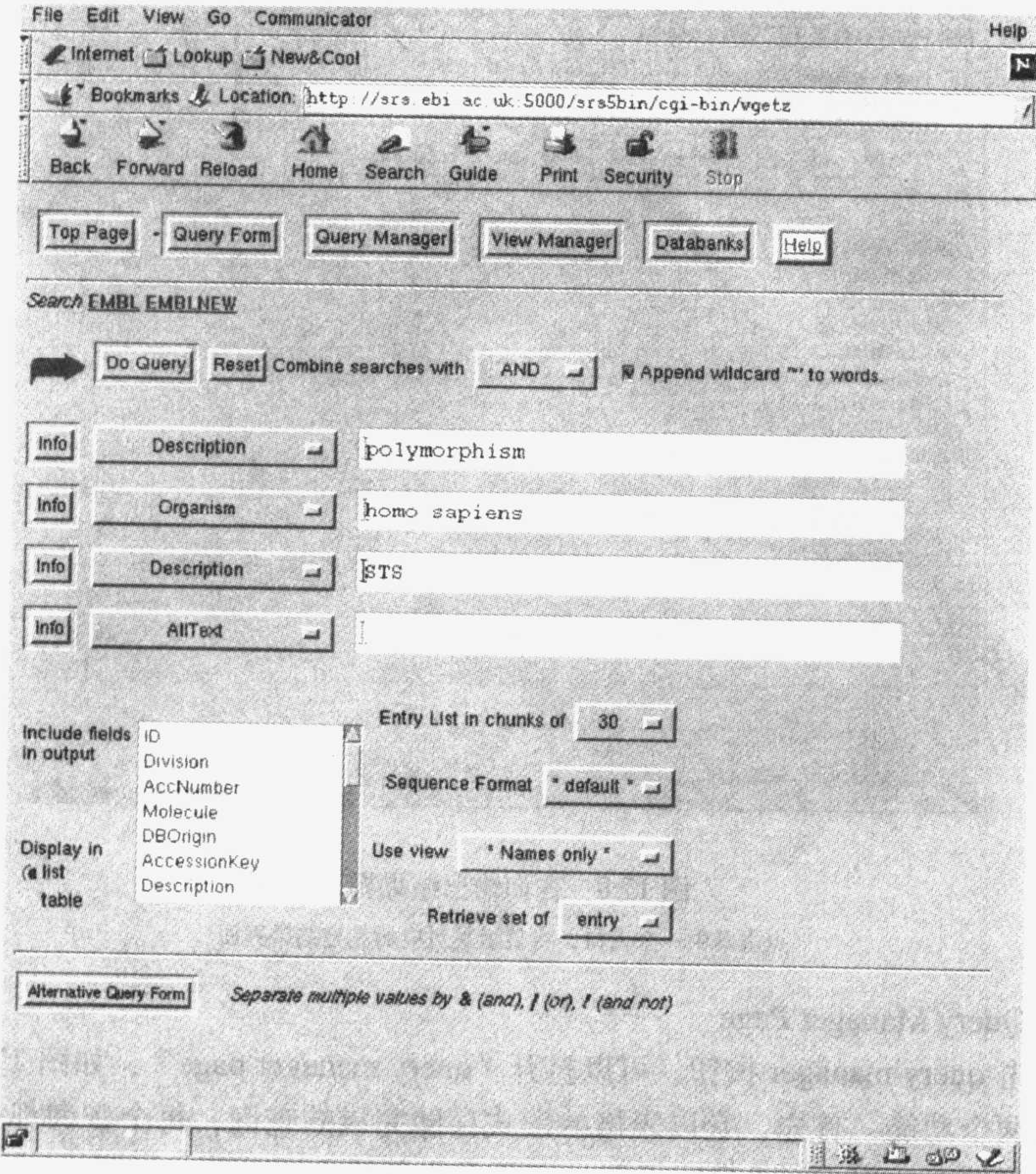


图 17.7 在查询界面运行的一个查询

3) 查询结果页

该页(图 17.8)显示提交查询的结果，显示了查询字符串、命中结果的数目和用复选框和所用字段(如在“查询表格页”中所选的)发现条目。查询到的条目根据搜索的数据库和条目识别号列出来，所列出的每个条目都和相应的数据库链接。点击条目，打开显示完整条目的“单个条目页”，与其他数据库的交叉引用以超级文本的形式链接到相关的条目上。要一起阅读多个条目，可以用复选框选中。本页上其他的选项是供高级用户用的，在 SRS 手册中有完整的描述。

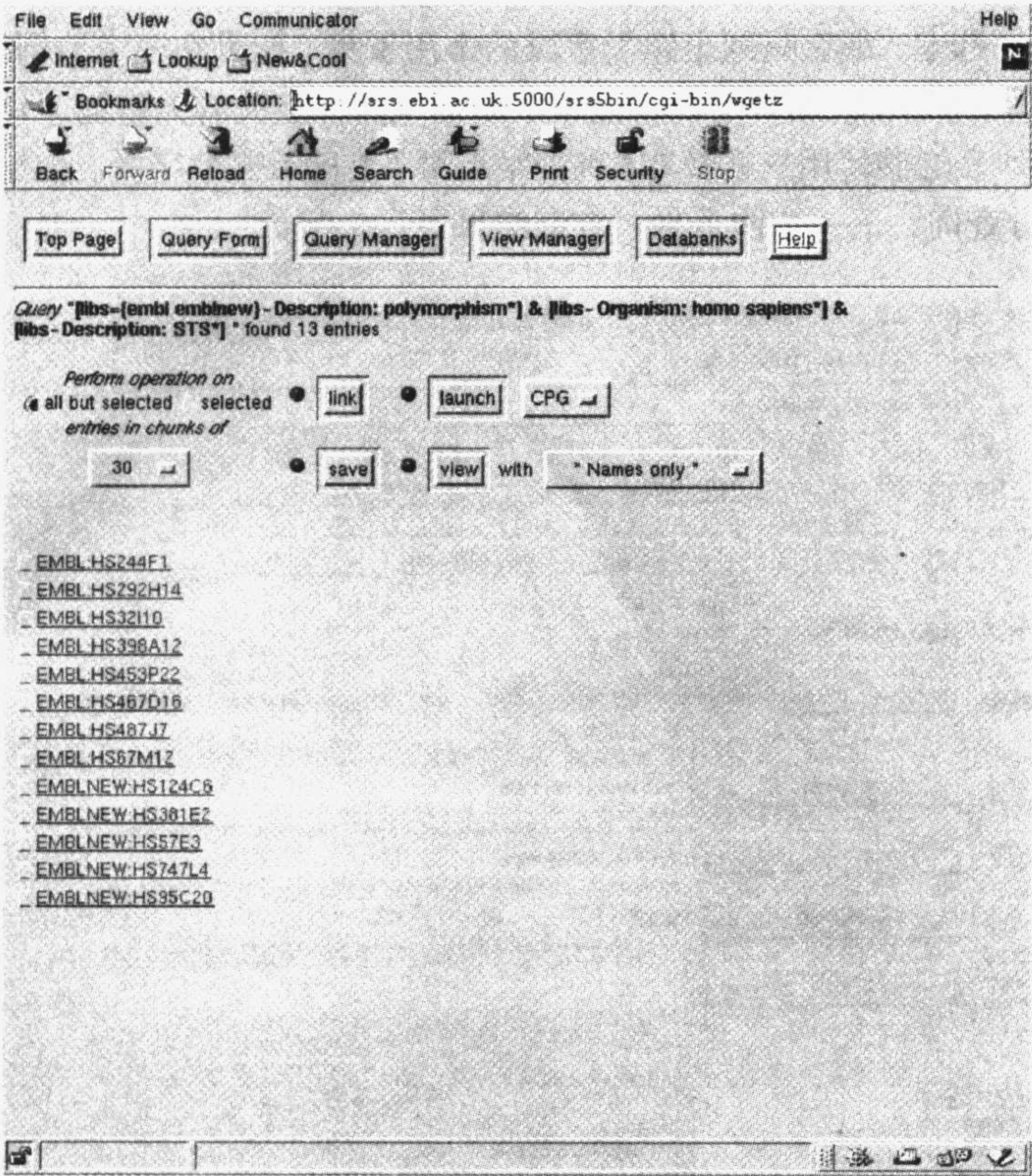


图 17.8 查询输出结果界面

点击其中一个入口，将显示其对应的完整结果界面

4) Query Manager Page

点击 query manager 按钮，可以打开“query manager page”，如图 17.9 所示。该页有两个功能，首先，它以表格的形式存储完成的查询，每个查询都会同时列出一个复选框，以 Qn(即 Q1、Q2、…)形式命名的查询名称，类型(即“select”、“query”、“link”、…)，检索到条目的总数，数据库名称，每个数据库的条目数，

在 SRS 查询语句中的查询表达方式和一个注释。“query manager”的第二个功能是进行进一步的查询和链接，具体在 SRS 手册中有说明。

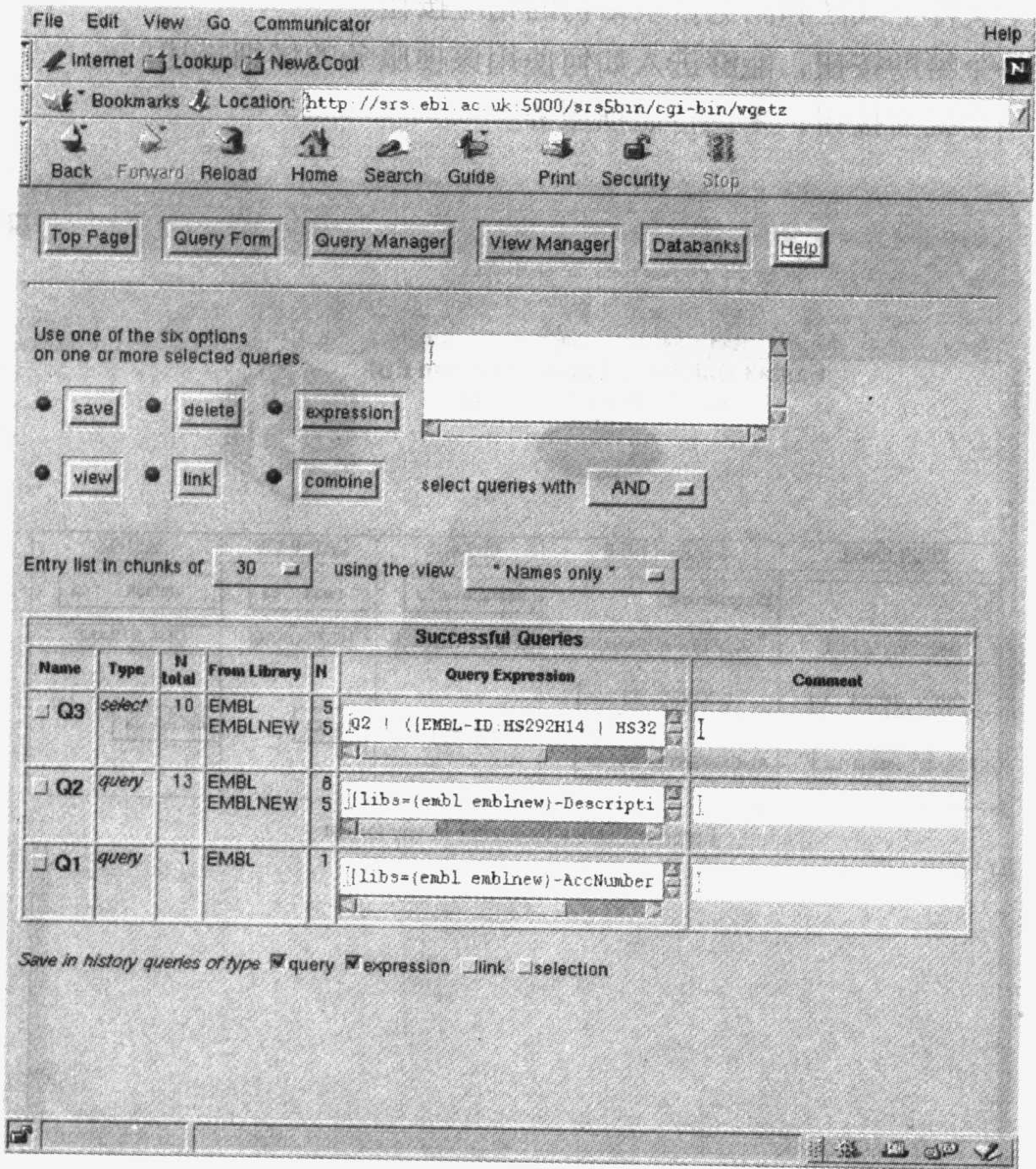


图 17.9　查询管理器显示了本当前所有的查询，它们可以被重用、组合和修改

17.3.3.2　序列分析工具

EBI 还提供大量的服务，使用户能够将自己的序列与在 EBI 数据库中可以得到最新数据进行比较。进入这些搜索工具的主要网页的地址为：http://www2.ebi.ac.uk/services.html。

该网页经常更新，来显示 EBI 现在可以提供的服务，所有的服务网页的外观基本相同(图 17.10)。

- 一个 e-mail 地址字段：取决于登陆的服务器或特定的查询参数，搜索以交互式或以批处理的形式进行。在后一种情况，结果将以 e-mail 的形式发送给用户。

- 输入查询特定参数的各种字段、复选框、菜单选择。
- 一个大的文本区域，可供用来剪切和粘贴或键入所要分析的序列，或一个文件上载区和浏览目录结构的相应按钮。
- 一个帮助按钮，能够进入如何使用该项服务的详细描述。
- 一个提交按钮，来执行该项服务。

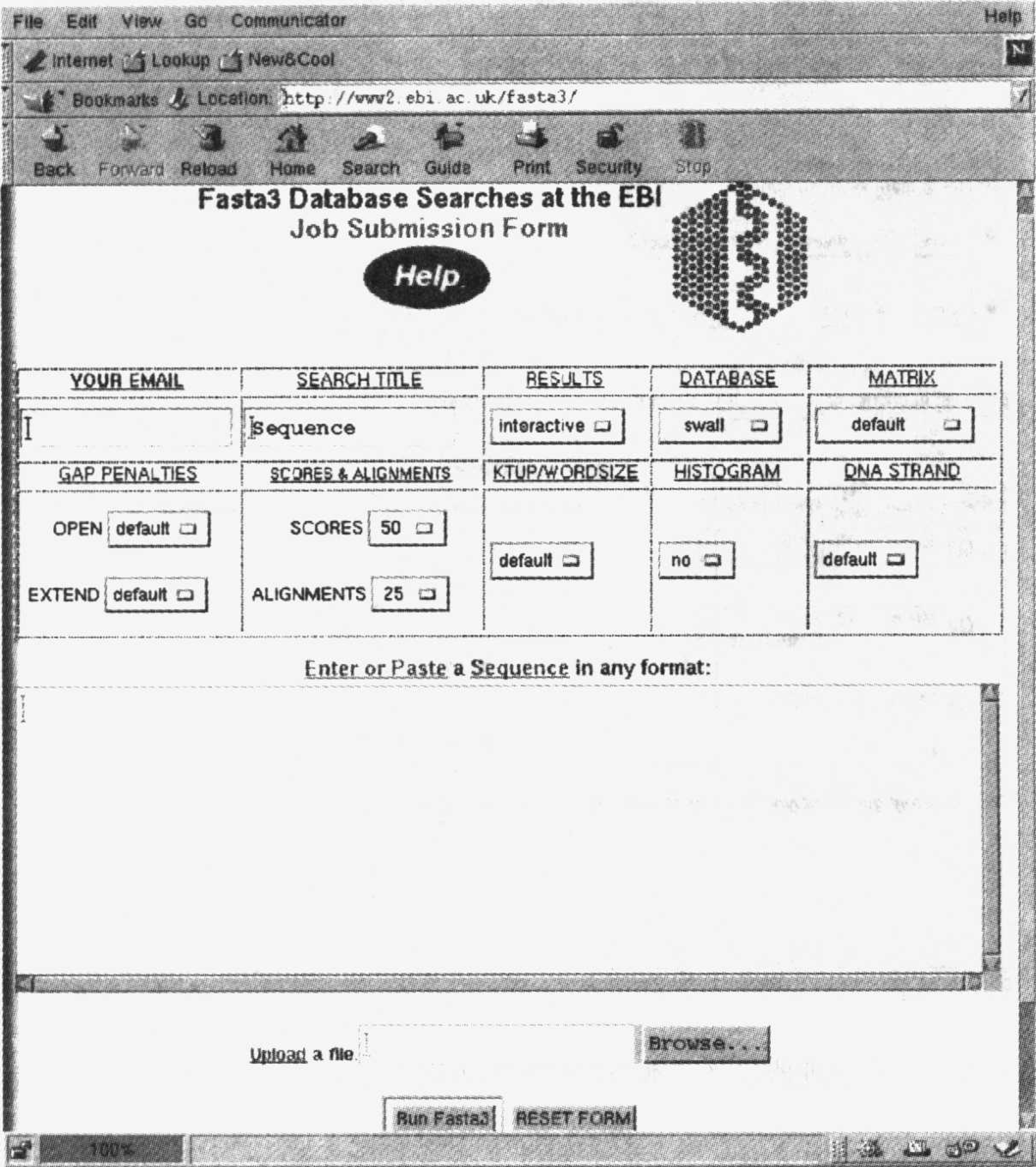


图 17.10 FASTA 服务页

作为分析页的一个实例：表格的顶部用于输入分析参数，文本区用于输入待分析的序列。在 Upload a file 文本字段键入文件名也可以从文件上载序列。页面的底部有两个按钮：一个按钮运行分析工具，一个是将表格重新设置为起始值

1) FASTA3

这是一个既可以以交互式服务也可以以 e-mail 服务的形式提供使用 W. Pearson Fasta[9, 10]相似性搜索算法进行序列搜索的软件。FASTA 工具广泛用于序列数据库相似性搜索、鉴定亲缘关系很远的 DNA 和蛋白质序列、鉴定序列的结

构相似性。可以在 http://www2.ebi.ac.uk/fasta3/得到。

2) BIC_SW

本服务能够实现在 Compugen's Bioccelerator(http://www2.ebi.ac.uk/bic_sw/)上运行的完全 Smith-Waterman 算法。

3) SCANPS

Scanps 由 G. Barton[11]开发，是一种蛋白质序列数据库的搜索工具。它将蛋白质或核酸序列与数据库中的序列进行比较，它使用各种动态编程算法(dynamic programming algorithms)，如 Smith-Waterman 的本地比对法。可以从 http://www2.ebi.ac.uk/scanps/进入。

4) BLAST

这是一个既可以以交互式服务也可以以 e-mail 服务的形式提供对蛋白质序列数据库进行搜索的工具。BLAST 算法的两种版本：NCBI[12]版本和华盛顿大学[13]版本都可以在 http://www2.ebi.ac.uk/blast2 处得到。Blast 通过一种算法提供快速的序列比对，在这种算法中，通过优化本地相似性的方法进行近似的比对。它广泛用于 DNA 和蛋白质序列的搜索、基因鉴定搜索和在长的 DNA 序列中对相似性的多个区域进行分析。

5) PPSEARCH

Prosite 数据库模式搜索，地址为：http://www2.ebi.ac.uk/ppsearch。

17.3.3.3 分析工具

EBI 上有一些分析工具，如下所述。

1) CLUSTAL W

CLUSTAL W[14]是一种序列的多序列比对工具，用户提供一系列序列，在 http://www2.ebi.ac.uk/clustalw 上可以得一个比对结果。

2) PRATT

PRATT[15]是一种能够使用户在一组蛋白质序列中搜索保守模式的工具，用户可以指定要搜索哪种类型的模式，还可以限定有多少个序列满足要报告的模式。该工具可以在 http://www.ebi.ac.uk/pratt 得到。

3) GeneMark

GeneMark[16]是一种基因预测服务，用来预测原核生物和真核生物的蛋白质编码区，它基于一种特殊类型的编码和非编码核酸序列的 Markov 链模型。可以从 http://www2.ebi.ac.uk/genemark 进入。

4) Dali 服务器

Dali[17]服务器(http://www2.ebi.ac.uk/dali)是一种比较三维结构的网络工具。用户提交一个拟查询的蛋白质结构的坐标，Dail 将其与 PDB 中的坐标进行比较，随后一个结构相似蛋白的多重比对结果被寄回。在理想的情况下，

通过三维结构比较可以揭示具有生物学意义的相似性，而这在序列比较中是得不到的。在 PDB 库中，与一种蛋白质具有结构相似性的蛋白质保存在 FSSP 数据库中，其地址为：http://www2.ebi.ac.uk/dali/fssp/fssp.html。

17.4 今后的发展和服务

分子生物学数据和基因组数据天生就非常复杂，而且进化得很快。不断地开发出新的算法，旧的也不断更新，随着对新技术的探索，EBI 的服务也将不断升级，其中的一项技术为 CORBA(Common Object Broker Architecture)，正在成为标准的技术，能够开发新的方法来更方便地进入数据库和服务。

17.5 如何与 EBI 联系

EBI 的邮政地址为：EMBL Outstation，the EBI，Wellcome Trust Genome Campus，Hinxton，Cambridge CB10 lSD，UK。

关于网络服务：

普通咨询：support@ebi.ac.uk

数据库咨询：datalib@ebi.ac.uk

RHdb 的提交和咨询：rhdb@ebi.ac.uk

BioCatalog 咨询：biocat@ebi.ac.uk

通过 e-mail 提交数据：datasubs@ebi.ac.uk

通过网络(EMBL)提交数据：http://www.ebi.ac.uk/submission/webin.html

通过 e-mail 纠正 EMBL 条目：update@ebi.ac.uk

通过网络纠正 EMBL 条目：http://www.ebi.ac.uk/ebi_docs/update.html

EBI 网络文件服务器：netserv@ebi.ac.uk

EBI ftp 服务器：ftp://ftp.ebi.ac.uk

EBI 万维网服务器：http://www.ebi.ac.uk

EBI 万维网搜索和分析服务：http://www2.ebi.ac.uk/Services

(马洪霞　张艳宇　译)

参 考 文 献

[1] Stoesser, G., Moseley, M. A., Sleep, J., McGowran, M., Garcia-Pastor, M., and Sterk, P. (1998) The EMBL nucleotide sequence database. *Nucleic Acids Res.* **26**, 8-15.

[2] Benson, D. A., Boguski, M. S., Lipman, D. J., Ostell, J., and Ouellette, F. (1998) GenBank. *Nucleic Acids Res.* **26,** 1-7.

[3] Tateno, Y., Fukami-Kobayashi, K., Miyazaki, S., Sugawara, H., and Gojobori, T. (1998) DNA data bank of Japan at work on genome sequence data. *Nucleic Acids Res.* **26,** 16-20.

[4] Bairoch, A. and Apweiler, R. (1998) The SWISS-PROT protein sequence databank and its supplement TrEMBL in 1998. *Nucleic Acids Res.* **26**, 38-42.

[5] Benrnstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. F. Jr., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T., and Tasumi, M. (1977) The Protein Data Bank: a computer based archival system file for macro-molecular structures. *J.Mol. Biol.* **112,**535-542.

[6] Lijnzaad, P., Helgesen, C., and Rodriguez-Tomé, P. (1998) The Radiation Hybrid Database. *Nucleic Acids Res.* **26,** 102-105.

[7] Rodriguez-Tomé, P. (1998) The BioCatalog. *BioInformatics* **14,** 469-470.

[8] Etzold, T. and Argos, P. (1993) SRS—an indexing and retrieval tool for flat file data libraries. *Comput. Appl. Biosci.* **9,** 49-57.

[9] Pearson, W. R. and Lipman, D. J. (1988) Improved tools for biological sequence analysis. *Proc. Natl. Acad. Sci.USA* **85,** 2444-2448.

[10] Pearson, W. R. (1990) Rapid and sensitive sequence comparis on with FASTP and FASTA. *Meth. Enzymol.* **183,** 63-98.

[11] Barton, G. J. (1997) *SCANPS version 2.3.11 User Guide.* University of Oxford, UK.

[12] Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990) Basic alignment search tool. *J. Mol. Biol.* **215,** 403-410.

[13] Altschul, S. F. and Gish, W. (1996) Local alignment statistics. *Meth. Enzymol.* **266,**460-480.

[14] Higgins, D., Thomson, J., and Gibson, T. J. (1994) ClustalW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22,** 4673-4680.

[15] Jonassen I. (1997) Efficient discovery of conserved patterns using a pattern graph. *Comput. Appl. Biosci.* **13,** 509-522.

[16] Borodvski, M. and McIninch, J. D. (1993) GeneMark: parallel gene recognition for both DNA strands. *Comp. Chem.* **17,** 123-133.

[17] Holm, L. and Sander, C. (1993) Protein structure comparison by alignment of distance matrices. *J. Mol. Biol.* **233,** 123-138.

18 计算机辅助分析转录调控区域: MatInspector 和其他程序

Thomas Werner

18.1 引言

通常计算机辅助的 DNA 序列分析是一项“本地专家”任务。但是，各种基因组测序项目和测序服务正在给科学家们提供越来越多的 DNA 序列数据。这些数据必须翻译成信息一起提交到实验计划中。

幸运的是，许多软件工具已经具备了该功能。在互联网上有许多此类程序。这样研究人员就无须在本地安装和维护所有的软件(这在非计算机专业实验室中并非易事)。基于浏览器的程序还能提供友好的用户界面。通常，这样只需要一个网页浏览器(如 Netscape 或 Internet Explorer)就够了，而浏览器程序在大多数的 PC 机和工作站上都有。所以，现在基于互联网的程序在分子生物学研究中很流行。本章将着重介绍这些工具。

基因组 DNA 分析可归纳成两类：第一类是确定蛋白质-编码区(即“基因查找”)；第二类是确定其他重要功能区域，如和转录调控有关的序列(启动子、增强子、LCR 等)。第二类分析是本章的重点，因为所介绍的软件都是用来分析转录调控的。

一般来说，启动子是基因的一部分，起介导和控制转录的作用，紧邻表达基因的上游。大多数蛋白质编码基因由 RNA 聚合酶 II 启动子控制，该启动子由数个不同功能的片段组成[1]。其中最重要的片段是蛋白质结合位点，如 TATA 框或其他蛋白质，如转录调控因子(TFs，如 AP-1，综述见参考文献[2])的结合位点。

尽管其他启动子元件，如自由卷曲 DNA、正向或反向重复元件也可能有助于启动子功能，但是蛋白质结合位点看来是最关键的元件。因此，检测 DNA 序列中可能的转录因子结合位点对所有启动子分析都非常重要。正如后面要详述的，分析启动子中这些结合位点的组织特征是完整的启动子分析的一部分。

18.1.1 IUPAC 同义序列和权值矩阵(weight matrix)

目前，有两个重要的方法进行 TF 结合位点识别：IUPAC 同义序列和 TF-结

合位点的权值矩阵。IUPAC 同义序列由频度最高的核苷酸组构成，这些核苷酸组(set)采用特殊的字母表示结合位点的某一位置上可能出现的核苷酸，而且不止一种，如 R 表示 A 或者 G。权值矩阵采用每一位置上核苷酸的全部组成来比对匹配序列，以获得一个更高复杂度的评价。例如，一个 12 序列比对中有一个位置是(T，T，T，T，T，T，T，A，A，A，C，C)(每一个字母代表每一个序列在该位置上的核苷酸)，那么在 IUPAC 同义序列中就指定为 T。新序列如果在此位置上是 T，那就认为是匹配序列，如果是 A 就会被忽略。即使是一个简单的核苷酸分布矩阵都会把 0.58 分配给 T，0.25 给 A，0.17 给 C。因此，权值分数代表测试序列和比对序列的相似程度，优于 IUPAC 同义序列。大多数基于权值的方法采用更复杂的加权函数，例如，通过比较实际的核苷酸分布和随机值或通过其他统计学检验(如信息含量，information content)。

18.1.2 权值矩阵优于 IUPAC 字串(string)吗？

IUPAC字串搜索程序在10年前就有了(如GCG软件包和SIGNAL SCAN[3])。但是，这些程序受它们源序列的影响很大，并且不能区分不显著的错配，即被结合蛋白耐受的错配和不能结合的序列的错配。另一方面权值矩阵对序列选择的敏感性较低，并且提供了一个定量的评分，提示蛋白质和所分析位点的相似程度。即使是在关键位置上的一个错配都会极大地降低匹配分值，并导致被正确地排除。

但是，采用权值矩阵分析 DNA 序列的 TF 结合位点需要预先编辑好的权值矩阵库，因为权值矩阵是不能由新手随意编辑的。只有权值矩阵搜索程序提供的预先编辑好的序列才能直接使用。MatInspector[4]就提供了这样的库，其他由生物计算以及生物信息学小组开发的程序，如 GSF-AG BIODV 的 FastM[5]和 Genom-Inspector[6, 7]都带有很好的库。

18.2 需要的资源

为分析未注释的 DNA 序列中的转录调控区域，需要准备几个软件工具，可分为两类：一类是分析单个序列的工具，另一类是分析一系列功能上相关序列的工具。

第一类 GSF-AG BIODV 开发的程序包括检测 TF 结合位点的程序，如 MatInspector 和 ConsInspector[8]，以及用来模建分析有条理的序列特征信息。这些程序包括 FastM 和 GenomeInspector。

第二类中有这样一些程序：MatInd、ConsInd 和 CoreSearch[9]，还有一个新的多序列比对程序 DIALIGN[10,11]，以及 ModelGenerator[12]，后者从一系列学习序列中开发出了有组织的模型。有关这些程序更详细的介绍和更多参考可在

GSF-BIODV 位于 http://www.gsf.de/biodv 的服务器上找到。MatInspector 和 FastM 都有网页界面，易于使用；其他程序需要下载并安装在本地的 UNIX 工作站上。下面要详细讨论 MatInd、MatInspector、FastM、ModelInspector 和 GenomeInspector。

18.2.1 MatInspector

MatInspector 是一个利用已预先编辑好的各种转录因子结合位点权值矩阵库来搜索序列中的相应因子可能结合位点的程序。程序为每一序列片段计算分值，采用所有的选定模型进行分析并报告所有的达标匹配。(MatInspector 支持矩阵组，真菌、昆虫、脊椎动物、植物和其他各种特定的矩阵选择。)模型相似性计算包括矩阵中单个位点的信息条目，这样能提高方法的敏感度(具体描述见参考文献[4])。一方面，如果在所有学习序列的矩阵中某一位点上只有 A，而测试序列是 T，那么该区域总体分值就会大幅降低。另一方面，如果矩阵中某一位点以几乎相等的机会出现 A、C 或 G，但没有 T，在测试序列中的 T 就不会对整体分值有太大的影响。

MatInspector 程序采用两套不同的相似性阈值，称为核心(core)和矩阵相似性(matrix similarity)。核心相似性指 4 个最保守的连续排列的核苷酸，通常表示蛋白质结合的最关键部位。在此区域的错配常常会影响蛋白质的结合，即使序列的其他部分匹配得很好。所以核相似性先作核对，以消除那些不太可能的蛋白质结合。这样会在大多数情况下减少伪匹配的数量，而具生物学意义的匹配会保留下来。

经验显示所包括的信息条目的核心相似性提高了 MatInspector 的相似性分值和实际的生物学结合亲和力的符合程度。这在一个最近的文献中得到了验证[13, 14]。

用全部的文库分析序列常常出现一大串的匹配结果，覆盖了整条序列。但是许多匹配都是无意义的。最主要的原因是缺省的阈值对某些矩阵来说太低(如 AP-1)，即使这样，还剔除了一些来自其他矩阵的有意义匹配。对于许多因子来说，有来自不同实验条件下的单独实验获得的模型。故不可能只使用一个模型而还能保留来自其他模型的全部有用信息。所以，MatInspector 得到的匹配单子会很长，使用者就要自己根据经验挑选最可能的匹配了。

多数(但不是全部)这样的任务可以通过 MatInspector 的商业化版本自动解决(Genomatix，http://genomatix.gsf.de)，它依靠一个扩展的模型库(比公共版本大)，这个库中的所有模型特征都经过矩阵相似性阈值优化(没有一般缺省值)。此外，相似和/或功能性相关的矩阵被合并到矩阵家族中，这样能充分利用所有矩阵的信息而无须在进行相关矩阵的冗余匹配。MatInspector professional 只报告每个矩阵家族的最优匹配。把优化的阈值和矩阵家族结合起来常常会把输出结果减

少到公用版本的一半，而不会显著降低实际的阳性匹配。但是，假阳性匹配还不能完全消除。用户还要考虑这些优势和花费是否相当。

需要注意的是，Genomatix 也在其服务器上提供一个增强的 MatInspector public 版本，供科学家免费使用。它能进行数据库搜索，所有工具都可以在这个网页上找到：http://genomatix.gsf.de/products。

18.2.2 FastM/ModelInspector

单一的 TF 结合位点不编码其自身的转录功能区(只结合蛋白)，因为生物学功能通常需要两个以上的 TF 结合位点。例如，TATA 框结合蛋白(TBP)有非常松弛的序列特异性。因此，原则上不可能为 TATA 框定义一个高特异性模型。并且每一个 TATA 框模型都产生很多明显看来不是 TATA 框的东西。然而，一个可能的 TATA 框(或其他 TF 结合位点)的周边序列常常对功能来说比结合位点的实际序列更重要。最简单的周边序列形式是一对结合位点。多对结合位点的协同效应已经为大家所熟知，并且还有一个这种功能对的数据库(COMPEL，参考文献[15])。因此，分析可能的启动子序列中与其他启动子相同的结合位点对是启动子识别过程中的第二步。

FastM 能使此类匹配模型容易产生，这些模型可用于扫描序列或 GenBank(或 EMBL)数据库，以查找其他序列中的类似匹配。这样 FastM 使我们对某一重要的匹配提出假设的模型并直接证实这些匹配是否有意义。可以扫描数据库中指定的结合位点发生反应信号，发现目的基因[5]。

真正的搜索工作是由 ModelInspector 完成的[12]，该程序找到和某一 TF 结合位点匹配的区域，并计算出全部模型的总分值。因为 ModelInspector 持续只报告高于完整模型阈值的匹配结果，所以假阳性数量通常比搜索单一匹配时要少得多。ModelInspector 的分值计算细节见参考文献[12]。

ModelInspector 和 FastM 都有商业版本(Genomatix 公司开发)。这些产品的亮点是不再局限于 TF 结合位点模型，因为这些模型的元件和模型本身含有 10 个不同的单独的元件，在公开版本只有两个。ModelInspector professional 还带有预先编译的功能性启动子库模块，用于直接扫描序列而无须先模建。

18.2.3 GenomeInspector

基于 FastM 的方法可以检测已经定义的功能性亚单位。GenomeInspector[6, 7] 是用来推导这些亚单位的，原理为仅仅从一个大片段的基因组序列(如酵母基因组)的重复序列特征来查找，而非原有的认识。

该分析的基本原理很简单。在一个线性分子，如 DNA 上的序列元件排列只有一个方法，即呈现一种序列次序(sequential order)(在一条链上一个接着一个排列)，而且还可以相隔一段距离。这就是 GenomeInspector 所要分析的：在一个给

定大小的窗口中元件对按相同的序列次序重复出现。

GenomeInspector 采用多个点和面集合进行相关分析。所有和某一转录因子结合位点模型匹配结果都表示为一个点集合，而更长的区段，如读框则表示为面集合。为了查找一个相关的 TF 结合位点匹配，GenomeInspector 分析相应的点集合，以发现异常的距离关联的位点匹配，每一个集合分析一种。例如，一种元件和酵母启动子的关联可以很容易地通过把结合位点模型的点集合与读框上游区域(由一个面集合表示)相联系来做到。这样，可以从原始序列中得到许多信息。这是该方法功能最强的地方，细节见参考文献[6]和[7]。

GenomeInspector 拥有一个基于 X-11 的图形用户界面，但是在互联网上无法访问。程序必须下载到安装了 X-11 操作系统的 UNIX 计算机上。GenomeInspector 是一个独立分析方法的大集合软件包，可看作是一个关联分析的工具箱。因此，程序功能多样化，用户需要培训，因为问题必须被翻译成 GenomeInspector 能够执行的流程。

18.3 方法

18.3.1 方法 1：权值矩阵的生成

如果在 MatInspector 的库中有 250 个以上的预编辑权值矩阵就可以略过本步骤。但是如果要创建新矩阵则另当别论。

(1) 登录 BIODV 的 WWW 服务器 `http://www.gsf.de/biodv`(或直接登录 MatInd/MatInspector 网页 `http://www.gsf.de/biodv/matinspector.html`)。该网页的下载选项在下半部分。下载并安装那些文件。

(2) 选择至少 4~5 个包含某一蛋白质结合序列。这些序列必须按 IG 格式保存在一个文件中：

命令行(可以多行)

序列编号(必须是不含空格的字符串)

AGCTGACGTCGACGTCG1(多块或多行序列，结尾必须是 1)

下一个序列(命令行，序列编号，序列)

(3) 启动 MatInd 程序，输入序列文件，让程序检测权值矩阵，如果数目达到了最小的质量要求(*re*<5)，就生成 3 个文件——2 个以扩展名为.sel 和.mat 的矩阵文件和一个以.ali 为扩展名的文件，后者含有具体的序列比对资料。如果随机期望值(*re*)大于 5，即在一个随机 DNA 序列的 1000 个核苷酸中期望多于 5 个匹配被找到，那么只生成.mat 和.ali 文件，表示矩阵质量低下。原因可能是包括的序列中不含结合位点，或者选择了不适当的序列区段(如在某些情况下结合位点被切断)。如果出现这种情况，训练序列就要改正(可能的话)并重新进

行 MatInd 分析。

(4) 新的矩阵文件现在可以拷贝到 MatInspector 的库目录下了，并且能够被本地安装的 MatInspector 使用。

也可以从一个预先编译好的核苷酸分布矩阵(无序列)中生成 MatInd 权值矩阵，这些模型常常在出版物中见到。请参考 MatInd 的用户手册(在下载的软件包中)。

注：MatInd 生成的自定义矩阵不能被 GSF 服务器的 WWW 版本 MatInspector (http://www.gsf.de/cgi-bin/matsearch.pl)采用。但是 Genomatix 提供了一个商业版本(MatInspector professional，http://genomatix.gsf.de)可以采用。

18.3.2 方法 2：用 MatInspector 分析序列

MatInspector 程序可以下载或在 WWW 服务器上使用。这里对 WWW 版本(http://www.gsf.de/cgi-bin/matsearch.pl)的使用方法进行介绍，本地安装的程序使用方法基本相同。

(1) 粘贴或上载欲分析的序列(上载和在目录中选择一个文件操作相同)。

(2) 设置搜索参数(阈值)。MatInspector 提供了核心和矩阵相似性的缺省设置。但这些值可以自行调节。除非有特别提示需要保持缺省值——它们在大多数的应用中都很有效。

(3) 从库中选择模型组分析序列。基本选项在网页的首页上，并且可以随后修改。MatInspector 提供了如下矩阵选项：

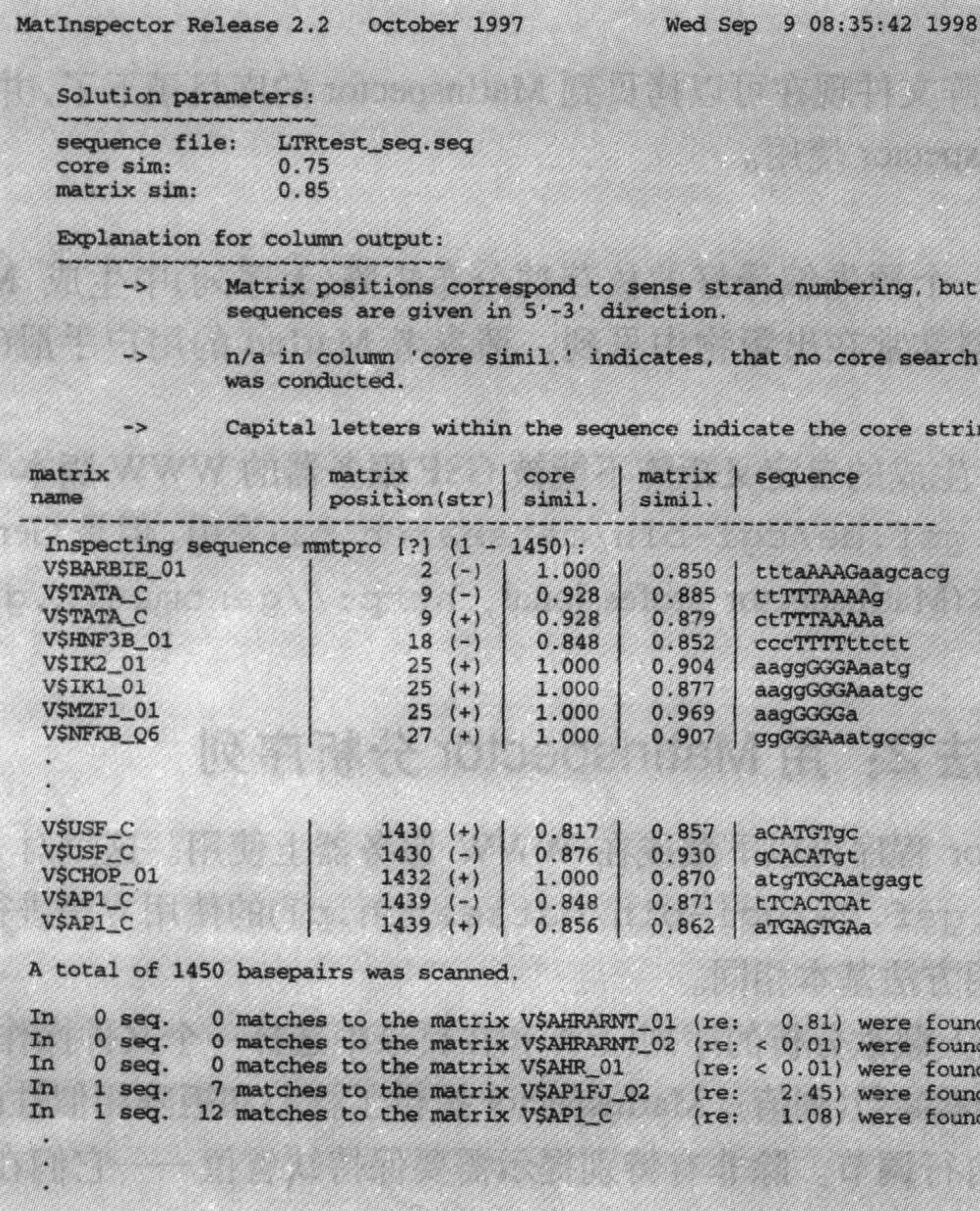
a. 选择整个文库(所有矩阵)。

b. 选择模型组的一个子集：脊椎动物、昆虫、真菌、植物、杂类模型(该子集的所有矩阵)。

c. 选择一个矩阵或数个矩阵(用户定义的模型集合)。

(4) 选择输出选项。MatInspector 允许用户按序列顺序、矩阵名称或匹配质量排列匹配结果。后者在通过提高矩阵阈值而不消除低分值位点而获得高分值位点时很有用。

(5) 提交查询。一旦选择了单独的模型(individual matrix selection)，这时会出现模型单子以供选择，只要在选定的模型复选框中单击即可。然后提交查询启动分析，只需数秒即可完成。结果立即以表格形式显示在网页中(图18.1)。各列分别表示矩阵名称，序列中的位置是匹配起始点，链的方向(+/–)，核心和矩阵相似性，以及发现匹配的具体序列。矩阵名称通常是一个超链接，然后更具体的矩阵描述在 TRANSFAC 数据库中可以找到[15]。在表的末尾提供了一个统计学总结，描述了每一个矩阵在序列中出现的频度，包括未找到的矩阵。



```
MatInspector Release 2.2   October 1997          Wed Sep  9 08:35:42 1998

    Solution parameters:
    ~~~~~~~~~~~~~~~~~~~~
    sequence file:   LTRtest_seq.seq
    core sim:        0.75
    matrix sim:      0.85

    Explanation for column output:
    ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
          ->      Matrix positions correspond to sense strand numbering, but all
                  sequences are given in 5'-3' direction.

          ->      n/a in column 'core simil.' indicates, that no core search
                  was conducted.

          ->      Capital letters within the sequence indicate the core string.

  matrix                 | matrix         | core   | matrix | sequence
  name                   | position(str)  | simil. | simil. |
 ------------------------------------------------------------------------
   Inspecting sequence mmtpro [?] (1 - 1450):
    V$BARBIE_01          |         2 (-)  | 1.000  | 0.850  | tttaAAAGaagcacg
    V$TATA_C             |         9 (-)  | 0.928  | 0.885  | ttTTTAAAAg
    V$TATA_C             |         9 (+)  | 0.928  | 0.879  | ctTTTAAAAa
    V$HNF3B_01           |        18 (-)  | 0.848  | 0.852  | cccTTTTttctt
    V$IK2_01             |        25 (+)  | 1.000  | 0.904  | aaggGGGAaatg
    V$IK1_01             |        25 (+)  | 1.000  | 0.877  | aaggGGGAaatgc
    V$MZF1_01            |        25 (+)  | 1.000  | 0.969  | aagGGGGa
    V$NFKB_Q6            |        27 (+)  | 1.000  | 0.907  | ggGGGAaatgccgc
    .
    .
    .
    V$USF_C              |      1430 (+)  | 0.817  | 0.857  | aCATGTgc
    V$USF_C              |      1430 (-)  | 0.876  | 0.930  | gCACATgt
    V$CHOP_01            |      1432 (+)  | 1.000  | 0.870  | atgTGCAatgagt
    V$AP1_C              |      1439 (-)  | 0.848  | 0.871  | tTCACTCAt
    V$AP1_C              |      1439 (+)  | 0.856  | 0.862  | aTGAGTGAa

  A total of 1450 basepairs was scanned.

  In   0 seq.   0 matches to the matrix V$AHRARNT_01  (re:    0.81) were found.
  In   0 seq.   0 matches to the matrix V$AHRARNT_02  (re:  < 0.01) were found.
  In   0 seq.   0 matches to the matrix V$AHR_01      (re:  < 0.01) were found.
  In   1 seq.   7 matches to the matrix V$AP1FJ_Q2    (re:    2.45) were found.
  In   1 seq.  12 matches to the matrix V$AP1_C       (re:    1.08) were found.
   .
   .
   .
```

图 18.1　MatInspector2.2 输出的分析一个长度为 1450 核苷酸的序列，该序列中包含有一个反病毒 B 类 LTR(长末端重复)。输出被截断了以便显示。中间 3 个点表示被省略的部分

18.3.3　方法 3：用 FASTM 分析序列组织

一旦可能的结合位点通过审查 MatInspector 的结果被确定，我们推荐用 FASTM 程序分析这些位点以找到可能的启动子内核，如 18.2.2 节所示。该程序也有万维网界面(http://www.gsf.de/cgi-bin/fastm.pl)，可以减小模建任务的复杂度，只要从一个万维网表格中选择即可。下面详述了如何核对可能的结合位点配对的有效性。

(1) 粘贴或上载欲分析的序列(上载和在目录中选择一个文件操作相同)。此外，还可以指定数据库类别，该类中所有序列都可以进行分析。

(2) 从表单中选择 TF 结合位点矩阵(和 MatInspector 文库的相同，这样程序发现的每一配对也都可以在 FASTM 中指定)。该步骤包括指定链的方向(sense/antisense)。除非矩阵完全对称，否则就有方向问题。也可以指定双向，每一个矩阵的核心和矩阵相似性在 FASTM 中是单独设置的(和 MatInspector 的搜索参数一致)。如果某一特定的结合位点矩阵不存在，就可以输入一个 IUPAC 同义

序列(consensus)。但是，基于 IUPAC 的模型特异性比基于矩阵的模型要差得多，只能在相当严格的 IUPAC 序列中使用(不能有 Ns!)。至于多个矩阵，IUPAC 序列链的方向可以指定或加入，FASTM 可以容忍 IUPAC 序列上的 1~2 个错配。

(3) 从第二个矩阵列表中选择下游 TF 结合位点的矩阵，或者指定一个 IUPAC 序列。链选择和错配选择和第一矩阵相同。

(4) 指定两个结合位点之间的距离范围。备选择项中提供了3 个常用的距离范围(close、medium 和 long)，并且还可以输入自定义的距离范围。距离范围越小，结果模型就更特异。

(5) 最后，要指定一个正确的 e-mail 地址以接收结果数据。这是必需的，因为 FASTM 的数据库类分析需要一些时间。

(6) 启动序列分析或数据库类别分析。结果文件如图 18.2 所示，这是一个搜索 TF 结合位点匹配的例子。搜索模型总结在表的上方，随后是每个单独的匹配细节。每一个匹配中，位置和各自矩阵分值都列出，并且还有模型的综合分值显示在结果的下面。最大的双矩阵模型分值可达到 2.0。

```
FastM / ModelInspector WWW Release 1.1              Wed Sep  9 08:43:16 1998

Sequence model of "seq":

      0.75/0.80 0.75/0.80

      V$GRE_C   V$GRE_C

      (+)       (+)

------[me]------[me]------


Distances between elements:

V$GRE_C          -          V$GRE_C:   80 -  100 bp (+/- 10 bp)

Input parameter:

Inspecting both strands of sequences(s).
Maximum number of matches in output file: 100

Inspecting sequence EP011026 (1 - 600):

(-)               ------------------------------------------------------------
                     position     core sim.    mat. sim.

    V$GRE_C               418 (-)       1.00          0.83
    V$GRE_C               308 (-)       0.86          0.85

element score =    1.67



Inspecting sequence EP023010 (1 - 600):

(+)               ------------------------------------------------------------
                     position     core sim.    mat. sim.

    V$GRE_C               264 (+)       1.00          0.80
    V$GRE_C               346 (+)       0.92          0.81

element score =    1.61

         .
         .
         .

Sequences searched: 1306 (783600 bp); 7 matches found in 7 sequences.
```

图 18.2 用 FASTM 程序通过一个距离为 80~100bp 的两个糖(肾上腺)皮质激素受体结合位点模型来分析真核生物启动子数据库的输出结果

0.75/0.80 是指模型中单个位点核心/矩阵相似性的阈值，(+)表示正义链方向，[me]表示单个元素是一个 MatInspector 分子质量矩阵。与矩阵匹配值分别以单个分值和总分显示

协同结合位点对的生物学功能通常在数据库中产生少于 100 个匹配，这时可以看到整个表单。一般地，匹配数目本身就是一个良好的特异性指标。通过这种方法，许多可能的结合位点匹配能很快地进行有效性测验，并添加更多有关分析序列中推定启动子的内部结构信息。

一个约 50 个预编译的启动子内核文库(大多数是结合位点对，全部都经过实验验证)可以从 Genomatix 公司的商业产品 MatInspector professional 中得到。该程序基本和 MatInspector 相同，但内核比单一矩阵的特异性高 1~3 个数量级，在输入一个数 10kb 的序列后可以把匹配数目降低到很少几个匹配。

基于 FASTM 的方法可以检测含有相同功能模块的序列组，即使缺乏整体的序列相似性。因此，这对常规的基于比对分析的工具，如 FASTA 或 BLAST 是一个扩展，这些工具严格地依赖与两个序列的整体相似性。如果其他属于功能上相关基因的启动子可以找到的话，一个模型搜索的结果会产生有关功能性启动子重要的信息。

18.4 结语

采用上述工具分析转录调控序列和几年前所谓的自动化测序有些相似。工具使几个重要的分析步骤更方便了。但是，正如测序仪，完全的成功只能由有经验的、熟知基本知识的用户达成。现在还没有可以使新手获得很好结果的方法。但是，如果用户对转录调控的原理很了解，这些工具就会有效地提高序列数据分析进程，并发现很多需要更多实验才能获得的信息。

万维网和浏览器导向的用户友好的界面开发克服了大多数该软件使用上的困难。对于用户来说，在没有自动化工具以得到满意结果的现在，学习如何把这些分析结果翻译出来是值得的。

致谢

该工作部分由德国联邦教育和研究部门(BMBF)资助，项目编号 FANGREB(Functional Annotation of Genomic Regulatory Regions) 0311641，以及欧盟资助项目 BI04-CT95-0226——转录数据库和分析工具(TRADAT)。

(邓景致 译)

参 考 文 献

[1] Smale, S. T. (1997)Transcription initiation from TATA-less promoters within eukaryotic protein-coding genes. *Biochim. Biophys. Acta—Gene Struct. Expression* **1351**, 73-88.

[2] Sauer, F. and Tjian, R. (1997) Mechanisms of transcriptional activation: differences and similarities between yeast, *Drosophila,* and man. *Curr. Opin. Genet. Develop.* **7**,176-181.

[3] Prestridge, D. S. (1996) SIGNAL SCAN 4.0: additional databases and sequence formats. *Comp. Appl. Biosci.* **12,** 157-160.

[4] Quandt, K., Frech, K., Karas, H., Wingender, E., and Werner, T. (1995) MatInd and MatInspector: new fast and versatile tools for detection of consensus matches in nucleotide sequence data. *Nucleic Acids Res.* **23,** 4878-4884.

[5] Lavorgna, G., Boncelli, E., Wagner, A., and Werner, T. (1998) Detection of potential target genes in silico? *Trends Genet.* **14,** 375-376.

[6] Quandt, K., Grote, K., and Werner, T. (1996) GenomeInspector: basic software tools for analysis of spatial correlations between genomic structures within megabase sequences. *Genomics* **33,** 301-304.

[7] Quandt, K., Grote, K., and Werner, T. (1996) GenomeInspector: a new approach to detect correlation patterns of elements on genomic sequences. *Comp. Appl. Biosci.* **12,** 405-413.

[8] Frech, K., Herrmann, G., and Werner, T. (1993) Computer-assisted prediction, classification, and delimitation of protein binding sites in nucleic acids. *Nucleic Acids Res.* **21,** 1655-1664.

[9] Wolfertstetter, F., Frech, K., Herrmann, G., and Werner, T. (1996) Identification of functional elements in unaligned nucleic acid sequences by a novel tuple search algorithm. *Comp. Appl. Biosci.* **12,** 71-80.

[10] Morgenstern, B., Dress, A., and Werner, T. (1996) Multiple DNA and protein sequence alignment based on segment-to-segment comparison. *Proc. Natl. Acad. Sci. USA* **93,** 12 098-12 103.

[11] Morgenstern, B., Frech, K., Dress, A., and Werner, T. (1998) DIALIGN: finding local similarities by multiple sequence alignment. *Bioinformatics* **14,** 290-294.

[12] Frech, K., Danescu-Mayer, J., and Werner, T. (1997) A novel method to develop highly specific models for regulatory units detects a new LTR in GenBank which contains a functional promoter. *J. Mol. Biol.* **270,** 674-687.

[13] Frech, K., Quandt, K., and Werner, T. (1997) Finding protein-binding sites in DNA sequences: the next generation. *Trends Biochem. Sci.* **22,** 103-104.

[14] Frech, K., Quandt, K., and Werner, T. (1997) Software for the analysis of DNA sequence elements of transcription. *Comp. Appl. Biosci.* **13,** 89-97.

[15] Heinemeyer, T., Wingender, E., Reuter, I., Hermjakob, H., Kel, A. E., Kel, O. V., Ignatieva, E. V., Ananko, E. A., Podkolodnaya, O. A., Kolpakov, F. A., Podkolodny, N. L., and Kolchanov, N. A. (1998) Databases on transcriptional regulation: TRANSFAC, TRRD and COMPEL. *Nucleic Acids Res.* **26,** 362-367.

19 计算机辅助的基因鉴定

Gautam B. Singh

19.1 引言

新基因的发现对医学有非常大的影响，因此遗传学的重要性日趋明显。人类基因组计划中多学科协作，其目的是鉴定基因组中的每个碱基。基因组中储存了合成各种蛋白质的蓝图，这些生物大分子使得生物在结构上和功能上可行。合成单个蛋白质的蓝图或程序称作基因，根据编码的蛋白质复杂程度不同，基因的DNA 序列的长度在 1×10^3~1×10^6bp 之间，高等真核生物有 30 000~40 000 个基因。估计基因组长度的 10%~20%编码蛋白质。基因鉴定的问题就是从这些未知的DNA 序列中识别出这些编码序列。

基因组早期的研究目标是构建物理图谱。然而现在则转向精细的序列测定。如此可以使我们研究这些长度为几十到几百 kb 的真核基因的结构和功能。真核基因中只有极少比例的长度编码蛋白质。起初使用传统的方法，如 cDNA 筛选、外显子捕捉、随机 cDNA 克隆，但长度超过数 kb 的基因研究则耗费非常大的劳动量。后来基因组测序中心通常使用计算机来进行外显子的预测，此外还使用一些其他的方法来检测基因。

自 20 世纪 80 年代以来，在理论生物学研究的基础上，有人就提出可以开发发现基因工具的软件。这些程序分析 DNA 的序列，基于局部密码子的使用频率、是否存在远古保守模式或者是否对随机序列而言有重要的偏离来标记一段可能的编码片段。在参考文献[1]和[2]中列出了一些基因预测的软件，同时概述了每种软件的优缺点。

通过对诸多疾病病因的研究，科学家知道了一些疾病的遗传基础，但是对使用计算机辅助的方式来发现新基因的需求也越来越甚。现在已经开发出来利用大量不同在理论和机械学习工艺上进行基因判断的软件。本章将对一些软件的使用做一概述，同时概述了在开发基因预测软件时使用的基因预测精确性的方法。

19.2 材料-基因鉴定系统

此节将对目前使用的在一段未知 DNA 片段中进行基因鉴定的软件做一概述。

19.2.1 AAT：分析和标注工具

AAT(Analysis and Annotation Tool，分析和标注工具)通过对蛋白质和 cDNA 文库中的序列进行比较来鉴定 DNA 序列中的基因。AAT 包括两对程序[3]：每一对程序都包括一个数据库搜索和一个比对子程序，第一对程序用来对拟查询(query)的序列和蛋白质数据库进行搜索，而第二对程序则是对 cDNA 数据库进行搜索。比对程序搜索数据库中所有序列的共有序列且将它们排列成一个多序列的比对方式来比较，来增强剪接位点的预见性。序列间相似之处较低的比对将被滤过，最终的蛋白质和对应 DNA 排列一起反馈给用户。

第一个程序对使用两个叫 DPS 和 NAP 的程序对需要查询的 DNA 序列和蛋白质数据库进行比较。DPS 程序用来查找拟查询的 DNA 序列中的一段序列和蛋白质数据库中的某种蛋白质间的高匹配性，而广泛使用的比对(alignment)程序 NAP 则是用来查找 DNA 序列和相匹配蛋白质序列间最佳的比对排列[4]。NAP 的比对模式包容了内含子和密码子间的移码，因此在使用 GT/AG 共有序列来鉴定剪接位点时可以精确确定内含子的位置。第二个程序对由 DDS 和 GAP 组成，用来对拟查询的 DNA 序列和 cDNA 文库数据库相比较。DDS 是在 BLASTN 程序的基础上加以改进[5]。GAP 程序功能强大，广泛用于有内含子的 DNA 序列和 cDNA 间的比对[6]。

AAT 的另外一个功能是对 DNA 序列提供自动注解帮助。传统上此项工作手工进行，通过 BLASTX 将一个 DNA 序列的编码区域和蛋白质数据间的比对建立起来并以 post hoc 的方式将该 DNA 序列和一个注解文件相链接。这种帮助可以提供一个基因可能具有功能的线索，而这种功能是相关的蛋白质序列所具有的。AAT 进行这种比对是基因预测的根本，另外 BLASTX 的排列结果容易存在移码突变错误，在 AAT 开发了一个 DNA-蛋白质序列间的比对程序来克服了 BLASTX 的这一缺点。

19.2.2 MZEF: Michael Zhang's Exon Finder(Michael Zhang 的外显子发现程序)

MZEF 是内部编码外显子预测程序，它使用二次方程判别分析(quadratic discriminant analysis，QDA)的方法来达到判断外显子和假外显子(pseudoexons)。这种方法是 HEXON[7](GeneFinder 中的 FGENEH 是 HEXON 的改进)中使用的，是早期的线性判别分析(linear discriminant analysis，LDA)中的统计学模式识别概念的延伸。在 QDA 的模式下，可以更精确地判断外显子和假外显子的分割界线(surface)。

其算法规则如下所述：分析每个满足 AG→ORF→GT 模板的可能的外显子，

满足最短长度标准的外显子是可能的外显子，必须将其与假外显子分开，这个可能的外显子用 9 个有价值的特征载体来表述，其组成有如下参数：外显子的长度、分支分值、供体两侧的 6 碱基偏爱频率和受体剪接位点之间各方面的差异等。这个 9 维的特征性的载体 χ 在以下源于 QDA 的 log 比率测试条件下被归为外显子或假外显子：

$$\eta=\log\left(\frac{p_1}{p_2}\right)=\log\left(\frac{p_1^0}{p_2^0}\right)-\left(\frac{\delta_1-\delta_2}{2}\right)-\left[\frac{1}{2}\log\frac{|\Sigma_1|}{|\Sigma_2|}\right] \tag{1}$$

在上述方程式中使用的参数为：μ_i 和 Σ_i 代表组平均数和从测试设置(测试设置由 1897 个真正的外显子和 184 217 个假外显子组成)中得到的协方差矩阵，p_1^0 和 p_2^0 的数值表示对于一个假定的外显子属于 G_1 真正外显子组还是 G_2 假外显子组优先的概率，$\delta_i=(\chi-\mu_i)^{\mathrm{T}}\Sigma_i^{-1}(\chi-\mu_i)$ 是观察到的特征载体 χ 和 μ_i 间的 Mahalanobis 距离的平方(squared Mahalanobis distance)，$|\Sigma_i|$ 是 Σ_i 协方差矩阵的值[8]。

19.2.3 GENSCAN

GENSCAN 起先是构建人类基因组序列的基因结构的概率模型(probabilistic model)，随后将此模型应用到基因的预测上来。基因的概率模型包括真核基因特殊的组成性和功能性单位，如外显子、内含子、剪接位点、启动子和 poly A 加尾信号。应用该模型的算法可以预测出这些单位和一个基因的一部分，而且使用 GENSCAN 做出的预测不仅反映了在蛋白质数据库中的基因类型，同时也是一个独立的估测，这种独立的估测是我们已知知识的补充信息。

GENSCAN 对 DNA 序列预测的模型基于 GHMM(generalized hidden Markov model)，它利用 DNA 双链特征，在 DNA 的一条链可能有多个基因，对 DNA 的另一条链和两条链也同样如此。一个简化的 HMM 示例如图 19.1 所示，弧线代表

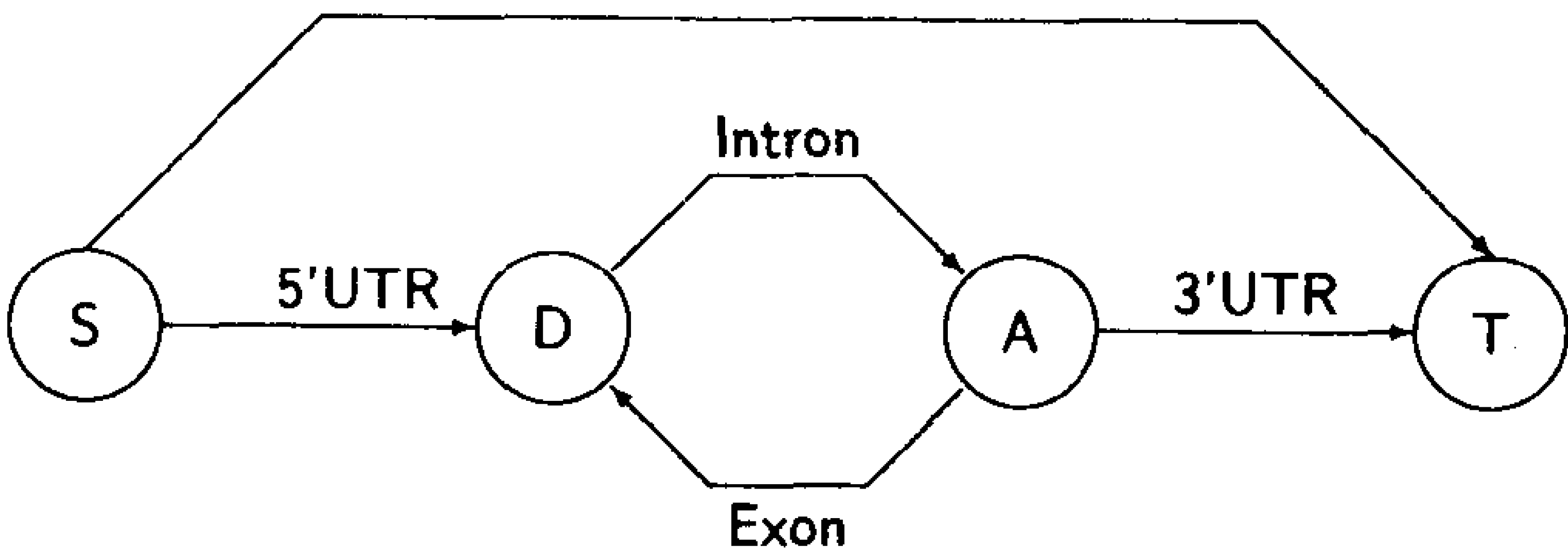


图 19.1 对于一个多外显子基因的 HMM，弧线代表了属于该基因功能单位的 DNA 序列，而节点代表了一个特殊的信号，该信号使得功能状态和分类序列从 5′→3′ 转换。节点 S 和 T 分别代表了起始和终止密码子，而 D 和 A 节点表示了供体和受体位点。从 S 到 T 的信号弧线使得 HMM 可以识别一个单个的外显子基因，该模型使用一组有特征基因的 DNA 序列可以不同状态信号转移的概率

了一组属于一类的核苷酸，节点代表了从一类核苷酸转变到另一类核苷酸的 DNA 区域。该程序使用最大依赖分解法(maximal dependence decomposition method)来模拟自然状态下编码蛋白的功能性信号转导及其相互间的作用的能力是器械性的(instrumental)，但是它是一般基因预测的强大工具。该程序的输出文本是一组预测到的基因和肽序列，而输出的画图是预测到的外显子的相对位置。该程序适用于脊椎动物、玉米和 *Arabidopsis* 序列的不同版本可以在参考文献[9]得到(脊椎动物版本也可适用于果蝇序列)。

19.2.4 Veil

VEIL(Viterbi Exon-Intron Locator)基于 HMMs(hidden Markov models)的观察，它为不同数据的模式序列提供了精确的概率法，因此，它使用自定义的 HMM 将非特征性的基因组 DNA 分割为外显子、内含子和基因间序列。exon-HMM 组件(module)可以获知在外显子中密码子使用的规律性以及外显子和除了符合读框的终止密码子的周期性。另一个相似的组件是 intron-HMM，在 HMM 模式下用类似管道的模式来描述可能的剪接位点，只是如管道的信号在长度上受到精确定义。例如，VEIL 的供体-受体位点模型分别由 9~15 个阶段组成。其他的 HMMs 包括起始密码子、进行 polyA 的信号化的 AAATAA，以及在起始密码子前的基因间序列组件。这些简单的模型综合在一起形成了和图 19.1 类似的一个总的基因模型。最终的 HMM 由 241 个阶段和 1003 个侧移(edge)组成。用 GENSCAN、VEIL 和 Genie 对同一个基因进行预测其结果大致相同，只是在个别片段有明显差异[10]。

在确定出这些区域后，运用 Viterbi 算法来分析需要查询序列的外显子和内含子。Viterbi 算法的基础是动态设计技术，在此算法中，确定了给定序列中期望要经历的最相似的一组状态，此外对于这个给定的序列所产生的或然性模型是计算机利用 Viterbi 算法(algorithm)计算过的，这代表了给定 DNA 序列包含一个基因的可能性。Glimmer 软件程序的目的和 HMM 相似，只是它是一种以内插值替换的(interpolated)HMM，且适用于微生物基因组的分析[11]。

19.2.5 Morgan

MORGAN(Multiframe Optimal Rule-Based Gene Analyzer)和其他基因发现软件的区别在于使用决策树(decision tree)分类技术，该技术源自统计学上的一致性(statistics community)，决策树在许多分类学广泛应用，如癌症的诊断、语音的理解、图像的理解和光学特征的识别。决策树经常应用于以其特征来表述的物体，如表述一个物体的 d-维(d-dimensional)特征可以表示为 f_1、f_2、…、f_d。后来这种分类过程植入树状结构，通过系列的试验，通常以 $f_k<T$ 的形式来鉴定未知物体。如此在决策树中的每个问题节点(question node)和一个线性判别相

对应，有助于在研究空间上将研究对象分割成一系列区室或者叶片，这些区室或者叶片的每一个个体都代表了一个我们在分类上感兴趣的整体。同样 MORGAN 在基因分类上存在两个分割(partitions)，分别用 C 表示编码，N 表示非编码。

MORGAN 系统的设计含 19 个特性设置(19-feature set)，这包括测量最长的可读框的特性、双密码子使用特性、6 碱基偏爱特性、密码子的使用特性、ACTG 位置上的不对称分布特性、2~9 周期的傅里叶系数(Fourier coefficients for periods 2~9)特性。决策树分析了 290 628 个样本，每个样本由 54bp 的人 DNA 序列组成。利用这些样本分析结果结合它们的预先分类，测试 MORGAN 决策树。从上述测试中得到了 20 个叶片和 19 个实验或者中间节点。对于一个给定的将被分为编码和非编码的 DNA 序列，首先对这 19 个特性(feature)的亚序列(subsequence)进行计算，最佳的分割依赖于随后的隔离赋值功能(separate scoring function)，给此分割赋予一个能反映此序列是一个外显子的可能性值，决策树给每个亚序列赋值，每个单独的赋值组合起来得出一个概率估计[12, 13]。

19.2.6 Genie

一个归纳隐藏马可夫模型(Generalized Hidden Markov Model)或 GHMM 在每个隐藏马可夫链(hidden Markov chain)状态下足以产生 DNA 序列的亚序列。在每个 GHMM 状态下产生亚序列的模型非常复杂，对另一个 HMM 也同样如此。在 Genie，每个组分都单独设计和测试，然后再组合形成一个模系统(modular system)，在测试设置中内含子和外显子的长度分布可以用来了解在 GHMM 状态下产生的字串的平均长度和可变性。供体和受体位点由神经网络识别，在该神经网络中供体位点为一个 15bp 的窗口，受体位点是一个 41bp 的窗口。这样就在图 19.1 中代表了供体和受体位点的节点(node)中植入了一个神经网络，对于一个给定的位置而言，这个神经网络可以返回这个位点是供体或受体的接续概率。在构建好这样一个模型后，通过测试就可以了解在 GHMM 状态下和对一个给定的特殊状态来产生每个核苷酸间转变的可能性。利用标准的基因数据设置，并应用机器学习技术来优化这些可能性(probability)[14]。

19.2.7 GeneFinder(FGENEH)

GeneFinder 是一组工具软件，可以从 Baylor 医学院获得，这些对人类基因组进行基因鉴定的软件包括：fgeneh 用来预测基因结构，fexh 用来预测 5′、内部序列、3′外显子，hspl 用来预测剪接位点，hexon 用来预测内部外显子。

fgeneh 的算法从预测所有可能潜在的内部外显子、潜在 5′和 3′外显子开始，使用一种线性判别式功能来描述每个外显子，这种线性判别式结合了这些外显子的前后特征，然后使用动态规划的方法来搜索这些外显子的最佳组合来

产生一个基因模型。在 hexon 中使用的算法同样基于可读框架的 GT 和 AG 侧翼序列判别性分析，线性判别式功能组合了供体和受体剪接位点、内含子 5′和 3′端区域的特征和可读框架的编码区域的特性。该软件只能预测在供体和受体剪接位点含有 GT 和 AG 的内含子，然而这通常包括超过 99%所有的可信的剪接位点。有理由期待该软件将来的版本会开发出包括其他物种的基因预测功能[7, 15]。

19.2.8 GeneParser

GeneParser 可以在给定的序列中找出特殊的特征，随后进行动态规划来得到符合其功能的最可能的构造，特别是在未知的 DNA 中搜索局部的剪接位点且估计，如密码子的使用、局部组成的复杂性、6-tuple 频率、长度分布和周期性的非对称性等符合的参数。该程序对给定序列之中所有亚内部序列中满足统计学上的可预测的，如内含子、外显子和它们的边界序列进行赋值。这些统计学上的赋值反馈给神经网络，由神经网络提供 log 可能性评价，这种评价将是对每个亚内部序列作出是否是内含子、第一个外显子、内部外显子和最后一个外显子的描述(这种输入反馈式的神经网络是经过优化的，可以预测出最多的正确结果)，然后应用动态规划算法对每个分类的亚内部序列进行计算来获得该基因模式总体最大的可能性。DP 算法的运行在内含子和外显子必须是毗邻的，而且是没有重叠的约束下进行。最高赋值组的结合将代表该基因模型最大可能性[16, 17]。

19.2.9 GeneLang

GeneLang 是一个循章办事式的识别系统，它使用计算机语言的工具和技巧来寻找基因和其他在生物数据库中高度有序的特征。例如，正式的计算机代码理论使用一套叫作语法的规则通过对给定的字母表来定义一组有效的字符串(string)，给基因制定语法的动机在于使得剖析器可以有效工作，如此一来剖析器就可以识别输入的字符串是否满足制定的基因语法的规则。这样在 GeneLang 软件中，使用上述的 DNA 语法在逻辑学上实现了程序化代码的 Prolog(programming language)[18]。该系统使用由 4 个 DNA 字母组成的文字来表述基因的供体和受体位点、内含子和外显子、起始和终止密码子等。如此一来基因就是有层次的，即该基因是由 ATCG 组成的句子，在 DNA 序列中的基因可以如英语语法上的正确的句子一样来阅读识别。

19.2.10 Grail

GRAIL 基因鉴定系统使用神经网络来识别基因，在这个神经网络中的 GRAIL 1、GRAIL 1a 和 GRAIL 2 将从许多编码预测器(coding predictor)中得到的数据组合

起来。至今，GRAIL1 的使用已经超过了 6 年，其神经网络识别在固定大小的100bp 的窗口内的编码能力，然后在不考虑其他的特征，如剪接的连接情况下计算该编码能力[19, 20]；GRAIL 1a 同样使用固定长度的窗口首先将潜在的编码区域定位，然后对每个潜在的编码区域周围诸多不连续的、不同长度的候选序列进行评价，对某个编码区域邻近的两个 60bp 序列得到的信息加以分析来发现这个编码区域正确的界限；在 GRAIL2，对每个潜在的可读框架使用不同长度的窗口，每个可读框架的边界为一对起始密码子/供体、受体/供体、受体/终止密码子，这样 GRAIL 2 利用基因组上的 context 信息对编码区域进行赋值，因此对于一个外显子附近无毗邻序列的序列而言是不合适的[21, 22]，然而就是这些改变提高了 GRAIL 2 的整体功能，特别是对于较短的外显子。GRAIL 1、GRAIL1a、GRAIL 2 都可以识别人类 DNA 序列上的编码区域，对于其他的生物体特别是哺乳动物它们仍然运行良好。

可以通过 Oak Ridge 国家实验室(ORNL)的 e-mail 服务等方式来获得 GRAIL，ORNL 对包含在电子邮件中的 DNA 序列进行处理，在 ORNL 也可以获得一个叫 Xgrail 的基于 X 的交互式的图表式的客户服务系统，Xgrail 支持包括基因模建和数据库搜索等分析软件工具。

19.2.11 基因鉴定的其他工具

GenView 系统使用双密码子统计来预测拼接信号和编码区域，该系统在分析 DNA 时的目标是将非真正的外显子数目最小化[23, 24]。GeneBuilder 同样使用双密码子统计来预测拼接信号和编码区域，使用动态规划操作(dynamic programming approach)[25]来构建基因可能的结构[26]。GeneID 系统是一个基于电子邮件服务的分析脊椎动物基因组 DNA、通过发现潜在的外显子来预测外显子和基因结构、使用一定的规则将这些因素组装成一个基因的软件[27]。PROCRUSTES 系统的算法基础是拼接比对算法(spliced alignment algorithm)，它搜索所有可能的外显子并选择和相关蛋白最符合的外显子作为基因的外显子[28]。GENMARK 系统使用 HMM 来预测外显子和内含子，从而定位编码区域[29]。

19.3 比较基因鉴定算法的方法

为了对基因发现系统进行比较，Burset 和 Guigo 定义了操作比较规则(performance comparison metric)，创建了数据组(dataset)[17, 30]，该数据组由 1993 年 1 月 GenBank 85.0 版本中的序列组成，这样，相对而言其组成是新的记录，该数据组的创建过程如下：

首先，Burset 和 Guigo 从 GenBank 收集了脊椎动物的蛋白质编码序列，然后通过一系列质控步骤，去除了含有假基因、符合读框的终止密码子、无内含子(特

别是 cDNA 序列)及存在不可理解的剪接连接(即在内含子的开始和终止无 GT 和 AG)的所有记录，此外他们还将免疫球蛋白和组织相容性复合物抗原的基因去除，最终得到了 570 个完整的序列，这 570 个序列分别为 570 个基因，每个基因至少有一个内含子，可从以下网址获得这些数据：

ftp://www-hgc.lbl.gov/pub/genesets/OtherDataSets/GUIGO_96

19.3.1 操作比较的参数

在核苷酸水平上进行的精确统计利用了下述参数。在序列中，如果被分析的某个核苷酸位于一个被预测为编码区域的序列中，那么将它归类为预测阳性(predicted positive，*PP*)，否则就是预测阴性(predicted negative，*PN*)，通过序列的标注可以得知核苷酸的事实阳性(actual positive，*AP*)或事实阴性(actual negative，*AN*)值，通过比较这些参数来计算真阳性(true positive，*TP*)、假阳性(false positive，*FP*)、真阴性(true negative，*TN*)和假阴性(false negative，*FN*)。预测程序(prediction program)的灵敏性(sensitivity, *Sn*)和特异性(specificity，*Sp*)在等式(2)和等式(3)有解释：

$$\text{灵敏性：}\quad Sn = \frac{TP}{AP} \tag{2}$$

$$\text{特异性：}\quad Sp = \frac{TP}{PP} \tag{3}$$

相似相关性(approximate correlation，*AC*)如等式(4)定义：

$$AC = \frac{\left(\frac{TP}{TP+FN}\right)+\left(\frac{TP}{TP+FP}\right)+\left(\frac{TN}{TN+FP}\right)+\left(\frac{TN}{TN+FN}\right)}{2} - 1 \tag{4}$$

在外显子水平，预测的外显子(predicted exons，*PE*)和有注解的外显子(annotated exons，*AE*)相比较。真外显子(true exons，*TE*)是预测的外显子的数目，这些外显子和有注解的外显子完全相同(即两个端点都是正确的)。外显子水平上的灵敏性和特异性在等式(5)和等式(6)分别给出：

$$\text{灵敏性：}\quad Sn = \frac{TE}{AE} \tag{5}$$

$$\text{特异性：}\quad Sp = \frac{TE}{PE} \tag{6}$$

在相关性测量(correlation measure)时，用 *Sn* 和 *Sp* 的平均值来测量在外显子水平上总体的精确度(accuracy)。在外显子水平上还要计算两个精确度(accuracy)：丢失外显子(missing exons, ME)，即没有被任何预测的外显子覆盖的有注解外显子的一部分；错误的外显子(wrong exons, WE)，即没有被任何真正的外显子覆盖的预测为外显子的一部分。由各个序列值的平均值来评价(calculate)一组序列的精确度测量(accuracy measure)，从而可以得到所有序列的平均值。

19.3.2 操作

表 19.1 中列出了上述的不同的工具软件，按照各个软件要正确地预测核苷酸时的 *Sn* 参数大小而排序，然而有时在预测外显子时比较(*Sn*+*Sp*)/2 更现实，在此情况下 MZEF、GENSCAN 和 AAT 最合适。

表 19.1 根据 Burset 和 Guigo[30]比较 570 个脊椎动物序列的数据组比较各种基因鉴定系统的精确性

方法	预测的核苷酸			预测的外显子		
	Sn	*Sp*	*AC*	*Sn*	*Sp*	(*Sn*+ *Sp*)/2
AAT	0.94	0.97	0.95	0.74	0.78	0.76
GENSCAN	0.93	0.93	0.91	0.78	0.81	0.80
MZEF	0.88	0.95	0.90	0.84	0.92	0.88
VEIL	0.83	0.72	0.73	0.53	0.49	0.51
MORGAN	0.82	0.80	0.78	0.58	0.54	0.56
Genie	0.78	0.84	0.77	0.61	0.64	0.63
GeneFinder	0.77	0.85	0.78	0.61	0.61	0.61
GeneID	0.63	0.81	0.67	0.44	0.45	0.45
GeneParser2	0.66	0.79	0.66	0.35	0.39	0.37
GeneLang	0.72	0.84	0.75	0.50	0.49	0.50
GRAIL-II	0.72	0.84	0.75	0.36	0.41	0.38
Xpound	0.61	0.82	0.68	0.15	0.17	0.16

19.4 结论

在大规模的人类基因组测序计划中，对于新得到的 DNA 序列中人类蛋白质编码基因的预测非常重要，将来会开发有更多的软件系统来增加在 DNA 中

进行基因鉴定时的可信度。虽然目前开发的软件系统的预测结果并非绝对准确，但是在目前高通量基因组测序计划时代，为了和序列数据的快速分析相适应，它们所提供的结果非常重要。目前所面临的问题是在预测内含子/外显子边界时精确度的不足反过来导致了在鉴定真核生物基因组结构时所得到的结果并不充分，以及对较短的外显子的处理能力不足，此外，大量预测上的假剪接位点最终降低了所预测外显子的可信性。由于不同的软件系统对 DNA 序列不同的加工处理过程，在以组合的方式来检查它们的结果时似乎是似是而非的。表 19.2 列举了由不同的基因鉴定软件系统所产生的编码的含共有序列的区域。而且，对于已知方法在方法学上应用的理解对于该方法所产生结果的解释至关重要。

表 19.2　万维网上基因鉴定网址表

方　法	站　点
AAT	http://genome.cs.mtu.edu/aat.html
GENSCAN	http://gnomic.stanford.edu/GENSCANW.html
MZEF	http://sciclio.cshl.org/genefinder/
VEIL	http://www.cs.jhu.edu/labs/compbio/veil.html
MORGAN	http://www.cs.jhu.edu/labs/compbio/morgan.html
Genie	http://www-hgc.lbl.gov/inf/genie.html
GeneFinder	http://dot.imgen.bcm.tmc.edu:9331/gene-finder/gf.html
GeneID	http://www.imim.es/GeneIdentification/ Geneid/geneid_input.html
GeneParser2	http://beagle.colorado.edu/~eesnyder/GeneParser.html
GeneLang	http://cbil.humgen.upenn.edu/~sdong/genlang.html
GRAIL-II	http://compbio.ornl.gov/Grail-bin/EmptyGrailForm
SORFIND	http://www.rabbithutch.com/
PROCRUSTES	http://www-hto.usc.edu/software/procrustes/index.html
GenView	http://www.itba.mi.cnr.it/webgene/

(张艳宇　张英霞　译)

参 考 文 献

[1] Singh, G. B. and Krawetz, S. A. (1994) Computer based EXON detection: an evaluation metric for comparison. *Intl. J. Genome Res.* **1**, 321-338.

[2] Fickett, J. (1996) Finding genes by the computer: the state of the art. *Trends Genet.* **12,** 316-320.

[3] Huang, X., Adams, M. D., Zhou, H., and Kerlavage, A. R. (1997) A tool for analyzing and annotating genomic sequences. *Genomics* **46,** 37-45.

[4] Huang, X. and Zhang, J. (1996) Methods for comparing a DNA sequence with a protein sequence. *Comput. Applic. Biosci.* **12,** 497-506.

[5] Altschul, S., Gish, W., Miller, W., and Myers, E. (1990) A basic local alignment search tool. *J. Mol. Biol.* **215,** 403-410.

[6] Huang, X. (1994) On global sequence alignment. *Comput. Applic. Biosci.* **10,** 227-235.

[7] Solovyev, V., Salamov, A., and Lawrence, C. (1994) The prediction of human exons by oligonucleotide composition and discriminant analysis of spliceable open reading frames, in *Proc. 2nd Intl. Conf. on Intelligent Systems in Molecular Biology,* Altman, R., Brutlag, D., Karp, R., Latrop, R., and Searls, D., eds.) AAAI Press, Menlo Park, CA, pp. 354-362.

[8] Zhang, M. (1997) Identification of protein coding regions in the human genome based on quadrati discriminant analysis. *Proc. Natl. Acad. Sci. USA* **94,** 565-568.

[9] Burge, C. and Karlin, S. (1997) Prediction of complete gene stuctures in human genomic DNA. *J. Mol. Biol.* **268,**78-94.

[10] Henderson, J., Salzberg, S., and Fasman, K. (1997) Finding genes in human DNA with a hidden Markov model. *J. Comp. Biol.* **4,** 127-141.

[11] Salzberg, S., Delcher, A., Kasif, S., and White, O. (1998) Microbial gene identification using interpolated markov models. *Nucleic Acid Res.* **26,** 544-548.

[12] Salzberg, S. (1995) Locating protein coding regions in human DNA using a decision tree algorithm. *J. Comp. Biol.* **2,** 473-485.

[13] Salzberg, S., Delcher, A., Fasman, K., and Henderson, J. (1997) A decision tree system for finding genes in DNA. Technical Report 1997-2003, Department of Computer Science, Johns Hopkins University, March 1997.

[14] Kulp, D., Haussler, D., Reese, M. G., and Eeckman, F. H. (1996) A generalized hidden Markov model for the recognition of human genes in DNA, in *Proc. 4th Conf. on Intelligent Systems in Molecular Biology,* June 1996. St. Louis, MO (States, D., Agarwal, P., Gaasterland, T., Hunter, L., Smith, R., eds.), AAAI Press, Menlo Park, CA.

[15] Solovyev, V., Salamov, A., and Lawrence, C. (1994) Predicting internal exons by oligonucleotide composition and discriminant analysis of spliceable open reading frames. *Nucleic Acid Res.* **22,** 5156-5163.

[16] Snyder, E. E. and Stormo, G. D. (1993) Identification of coding regions in genomic DNA sequences: an application of dynamic programming and neural networks. *Nucleic Acid Res.* **21,** 607-613.

[17] Snyder, E. E. and Stormo, G. D. (1995) Identification of coding regions in genomic DNA. *J. Mol. Biol.* **248,** 1-18.

[18] Dong, S. and Searls, D. B. (1994) Gene structure prediction by linguistic methods. *Genomics* **23,** 540-551.

[19] Uberbacher, E. and Mural, R. (1991) Locating protein coding regions in human DNA sequences using a multiple-sensor neural network approach. *Proc. Natl. Acad. Sci. USA* **88,** 11 261-11 265.

[20] Uberbacher, E. and Mural, R. (1991) GRAIL seeks out genes buried in DNA sequence. *Science* **254,** 805.

[21] Uberbacher, E., Xu, Y., and Mural, R. (1996) Discovering and understanding genes in human DNA sequence using GRAIL. *Comp. Meth. Macromol. Seq. Anal.* **266,** 259-281.

[22] Xu, Y. and Uberbacher, E. (1997) Automated gene identification in large-scale genomic sequences. *J. Comp. Biol.* **4,** 325-338.

[23] Milanesi, L., Kolchanov, N., Rogozin, I., Ischenko, I., Kel, A., Orlov, Y., Ponomarehko, M., and Vezzoni, P. (1993) GenView: a computing tool for protein-coding regions prediction in nucleotide sequences, in *Proc. 2nd. Intl. Conf. on Bioinformatics, Supercomput. and Complex Genome Analysis* (Lim, N., Fickett, **J.,** Cantor, C., and Robbins, R. J., eds.) World Scientific Publishing, Singapore, pp. 573-588.

[24] Milanesi, L., Kolchanov, N., Rogozin, I., Kel, A., and Titov, I.(1993) Sequence functional inference, in *Guide to Human Genome Computing* (Bishop, M. J., ed.) Academic, New York, pp. 249-312.

[25] Rogozin, I. B., Milanesi, L., and Kolchanov, N. A. (1996) Gene structure prediction using information on homologous protein sequence. *Comput. Applic. Biosci.* **12,**161-170.

[26] Thomas, A. and Skolnick, M. (1994) A probabilistic model for detecting coding regions in DNA sequences. *IMA J. Math. Appl. Med. Biol.* **11,** 149-168.

[27] Guigo, R., Knudsen, S., Drake, N., and Smith, T. (1992) Prediction of gene structure. *J. Mol. Biol.* **226,** 141-157.

[28] Gelfand, M., Mironov, A., and Pevzner, P. (1996) Gene recognition via spliced sequence alignment. *Proc. Natl. Acad. Sci. USA* **93,** 9061-9066.

[29] Borodovsky, M. and McIninch, J. (1993) GENMARK: parallel gene recognition for both DNA strands. *Comp. Chem.* **17,** 123-133.

[30] Burset, M. and Guigo, R. (1996) Evalution of gene structure prediction programs. *Genomics* **34,** 353-367.

20 万维网上适用于一般用户和生物学工作者的 Primer3 程序

Steve Rozen　Helen Skaletsky

20.1 引言

设计 PCR 和测序用的引物是所有分子生物学工作者的一项重要工作。本章所涉及的 PCR 原理和操作参看文章后的参考文献[1]~[4]。

Primer3 为适应各种不同需要而提供 PCR 引物的计算机程序，例如，为放射杂交图谱而建立 STSs(序列标记位点)[5]，或者为单核苷酸多样性发现而扩增序列[6]。Primer3 也能为测序反应选择单引物及设计寡核苷酸杂交的探针。

在为引物和杂交探针选择寡核苷酸时，Primer3 能考虑到诸多因素。这些因素包括寡核苷酸熔链温度、片段长度、GC 含量、3′端的稳定性；估计的二级结构，扩增不想要片段(如散在的重复序列)的合适的退火温度，相同引物两个拷贝之间引物二聚体形成的可能性，以及源序列的准确度。在设计引物对时，Primer3 还考虑到产物的大小和熔链温度，两个配对引物之间引物二聚体形成的可能性，引物之间熔链温度的差异，以及相对于感兴趣或欲避免的特定的区域中引物的定位。

20.1.1 通过 Primer3 的万维网界面或作为一软件组成使用 Primer3 程序

大多数临时用户往往喜欢通过 Primer3 的万维网界面来使用该程序(图 20.1)，这适合于从少数序列中挑选引物。该界面的详细讨论请参阅 20.2 节。

要从成千上万条序列中挑选引物的科学工作者喜欢将 Primer3(尤其是 primer3_core 程序)作为一软件组成来使用，它能接受其他程序以方便的格式产生的输入，以及生成其他程序以方便的格式解释输出信息(我们认为没有人愿意人工处理成千上万条序列挑选的引物结果)。我们在 20.3 节中举例介绍了 primer3_core 作为软件组成的使用。下面的引物设计过程对万维网界面和 primer3_core 是相同的，事实上，万维网界面用的是 primer3_core 顶层所用的万维网 CGI 协议[9, 10]。

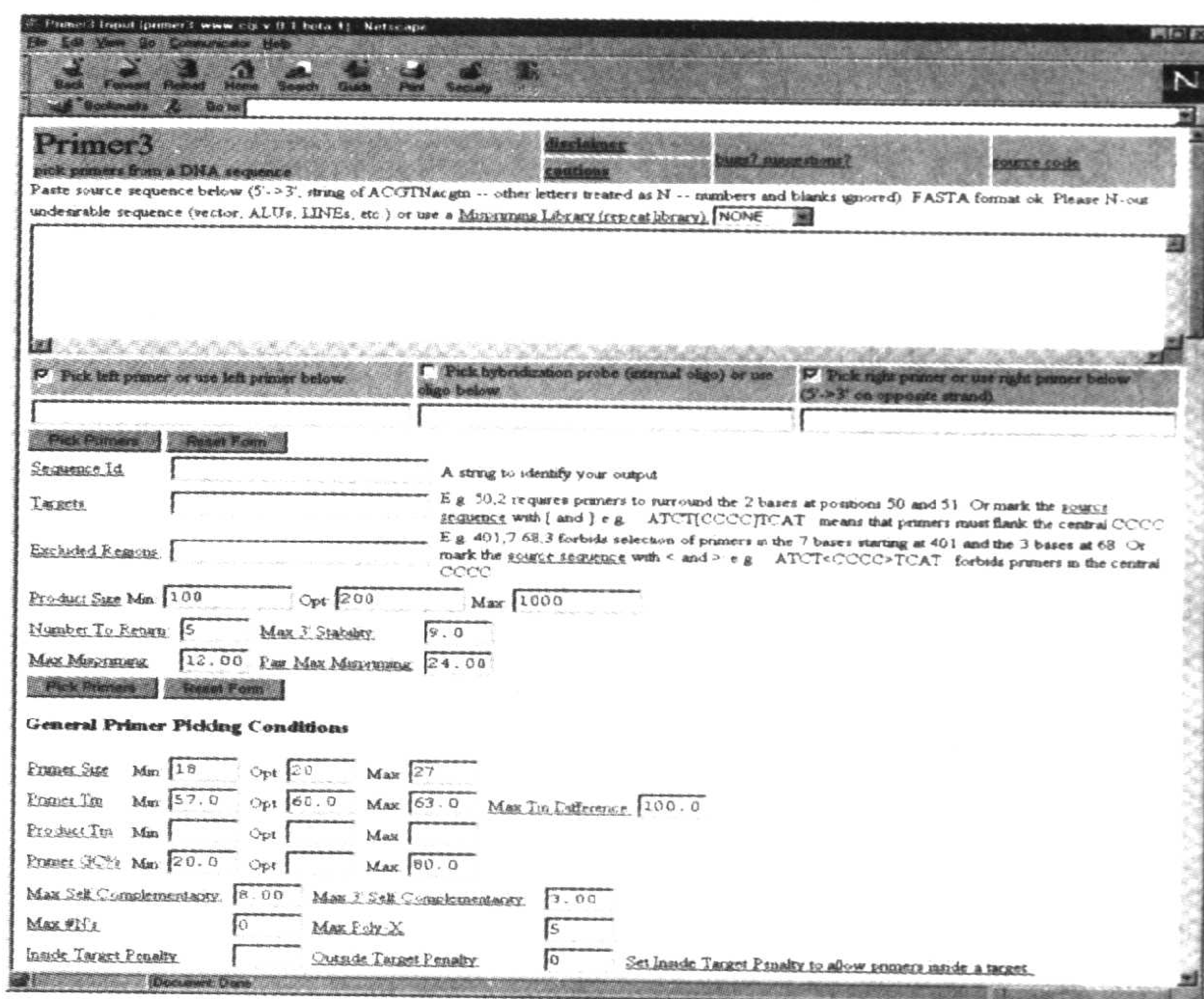


图 20.1　Primer3 的输入界面(用户未输入)

还有少数人喜欢对 primer3_core 进行修改而并不简单地把它作为一个软件组成来使用。因此，为最大限度的方便性和可修改性，我们用标准 ANSI C 语言[7]来编写 primer3_core 程序，使用标准 POSIX[8]并调用简单、通用的 ASCII 码输入和输出。此外，该程序还包括一套彻底的 primer3_core 检测系统，它能较容易地确保修改不会掺入错误。

尽管万维网界面的自解释程序对许多临时用户来说已足够了，但我们这里介绍的背景材料对其他人将会有帮助的。对一些有潜能的使用者、用户及分子生物学程序员，本章也介绍了使用 primer3_core 对手边特定的引物挑选任务去编写程序。

20.1.2　何处查找 Primer3

用户可以通过任何一个 Web 浏览器(如 Netscape)从 `http://www.genome.wwi.mit.edu/cgi-bin/primer/info.cgi` 网址进入公共万维网界面。用户也可从该网址下载 Primer3 程序。primer3_core 程序仅以资源的形式被利用，并且产生一个由 C 语言编写的可执行程序，详细参阅 20.3.1 节。万维网界面的源代码是可利用的，在计算机上运行 Web 服务器来使用它。可以对万维网界面(像 primer3_core)进行修改以满足一些特殊终端用户的要求。它是用这种应用的 perl 语言来编写的[11]。

20.1.3 Primer3 不能提供哪些帮助

很遗憾我们没有资源发布 Primer3 为即用型可运行形式，即没有把 Primer3 做成带有本地机客户端的(如 Microsoft Windows 或 Mac)形式，也没有以磁带、磁盘及 CD 盘形式让用户使用。其他的引物选择软件以完全支持的商业化形式(尽管它可能并不作为商业化软件组成)被利用。包括 OLIGO®的范例可通过以下网址或形式获得：Molecular Biology Insights，Inc.，Cascade，Colorado，http://www.mbinsights.com/[12]，LaserGene 的 DNAStar's PrimerSelect 模块(http://www.dnastar.com/)，以及遗传学计算机小组的 Wisconsin 软件包的 Primer(引物)模块。引物挑选程序包括 Primer 0.5[13](Primer3 以它为基础，但它可作为一独立程序及 Mac 和 PC 可执行的就绪运行形式)和 OSP(oligonucleotide selection program)[14]，可从院校机构得到。

Primer3 不能完成以下的任务：

- 给每一个引物自动添加标准 5′尾部。
- 挑选嵌套引物对。
- 为多重扩增挑选引物。
- 在一个序列上设计一系列扩增子。
- 从逆翻译氨基酸序列中挑选引物。

(然而，我们已将 primer3_core 作为一软件组成使用，与其他的代码连接来完成以上除最后一条的各项任务。)上面所提及的软件包能执行其中的一些任务。

20.1.4 PCR 和引物设计的应用是多种多样的

引物设计确实存在许多不同的问题。有时有人希望为大量序列设计的引物，或如果因为某些原因很难为一特定的源序列寻找好的引物时人们只能简单地丢弃源序列。下面将以高通量全基因组图谱(Primer3 及其前驱原本就是为它而设计应用的)为例来说明。在这个例子中人们要从成千条序列的数十条中设计 STSs，然后用这些 STSs 去上百次扩增。在这种应用中，与引物和亚序列扩增的成本相比，没有一个序列是有特殊价值的，因此只要某一序列的引物对可疑或不确定，就不值得再继续进行下去了。

在另外一些应用中人们想要设计引物尽可能去扩增一特定的序列；假如在没有明显的好的引物时，他会选择几种可能性的引物希望其中至少有一个能用。这样的例子包括，设计引物去区分两个非常相似的序列，或通过很难找到好的引物的 CpG 岛去扩增特定的外显子侧翼序列。在这种状况下最珍贵的资源就是特定序列的扩增，科学工作者将会花费很大的努力去得到一个纯的扩增子。

在引物设计方面还有其他各种各样的目的。有时有人想得到比较大的扩增子(如尽可能去扩增很多的 cDNA)，也有时还想得到比较短的扩增子(如尽可能接近

的 flank 一单核苷酸多形态)。有时扩增模板很复杂(如哺乳动物的基因组)，有时又很简单(如一个人工的单细菌染色体)。一些 Taq 的形成比引物二聚体或自身引发发夹结构的产生要少。

由于引物设计确实存在各种各样的问题，Primer3 能提供给用户大量的任选项去确定哪一种引物是可以接受的，以及哪一种引物比别的更好。这样大的选择量对少数经验丰富的用户也是无法克服，但对少数需要对设定值进行改变的特殊应用是例外。

本章不能讨论所有的 Primer3 的任选项，但却涉及了那些你最想要修改的部分。万维网界面和 README 用程序文件呈献给大家更多独特的任选项。

20.2 Primer3 在终端用户应用前景

这个章节主要介绍万维网界面，连同 primer3_core 一起演示几乎所有的引物挑选工作。

Primer3 执行输入一条序列和挑选单引物或 PCR 引物对。图 20.2 举例介绍在万维网界面的输入。用户已把源序列粘贴到该页顶端附近的大的数据输入区，然后选择 RODENT Mispriming Library 并进入 Sequence Id(“Sequence 1 为例”)和 Target(“40、78”)。

图 20.2　用户输入序列后的 Primer3 界面

源序列区的下面是标记了“Pick left primer or use left primer below”,“Pick hybridization probe(internal oligo) or use oligo below”和“Pick right primer or use right primer below”三个复选框。这些复选框支配 Primer3 是否想要设计一个引物对，或一个引物对加探针,或一格单独引物(如为测序用)或杂交探针。在图 20.2 中“Pick left…”和“Pick right…”框选中后，Primer3 就开始选择 PCR 引物对。

三个复选框的每个下面都是一个输入框。在这三个框中的其中一个放置一寡核苷酸序列便可指导 Primer3 去评价寡核苷酸并选择匹配的寡核苷酸(依赖于复选框)。

用户敲击输入页(图 20.2)Pick Primers 的任意一个键后 Primer3 返回便提示如图 20.3。20.2.2 节所示的供选引物并详细讨论这种输出的解释；当 Primer3 无法找到任何可接受的引物或引物对时，20.2.3 节还提出了一些继续进行的策略。

```
PRIMER PICKING RESULTS FOR Example Sequence 1

Using mispriming library rodent_ref.fasta
Using 1-based sequence positions
OLIGO            start  len      tm     gc%   any    3'   rep seq
LEFT PRIMER         19   20   59.06   50.00  2.00  0.00 12.00 ttcatctggttcccttcctc
RIGHT PRIMER       175   20   60.42   55.00  2.00  0.00 12.00 ggggtcaaggggaataacac
SEQUENCE SIZE: 180
INCLUDED REGION SIZE: 180

PRODUCT SIZE: 157, PAIR ANY COMPL: 5.00, PAIR 3' COMPL: 2.00
TARGETS (start, len)*: 40,78

    1 gcaacagtgaagtttctttcatctggttcccttcctccctaggcattggcaagcataag
                        >>>>>>>>>>>>>>>>>>>> *********************

   61 tcatgtggcccatgtcactattacaaccaaaacagctgatgggaaatgtgcgtacaggta
      *********************************************************

  121 tgtgataatgacctctgtgtccaccctaaacatacgtgttattccccttgacccccgcta
                                        <<<<<<<<<<<<<<<<<<<<

KEYS (in order of precedence):
****** target
>>>>>> left primer
<<<<<< right primer

ADDITIONAL OLIGOS
                     start  len      tm     gc%   any    3'   rep seq

 1 LEFT PRIMER          16   21   60.43   47.62  2.00  0.00 11.00 tctttcatctggttcccttcc
   RIGHT PRIMER        175   20   60.42   55.00  2.00  0.00 12.00 ggggtcaaggggaataacac
   PRODUCT SIZE: 160, PAIR ANY COMPL: 5.00, PAIR 3' COMPL: 2.00

 2 LEFT PRIMER          17   20   58.57   50.00  2.00  0.00 10.00 ctttcatctggttcccttcc
   RIGHT PRIMER        175   20   60.42   55.00  2.00  0.00 12.00 ggggtcaaggggaataacac
   PRODUCT SIZE: 159, PAIR ANY COMPL: 5.00, PAIR 3' COMPL: 2.00

 3 LEFT PRIMER          17   21   59.56   47.62  2.00  0.00 11.00 ctttcatctggttcccttcct
   RIGHT PRIMER        175   20   60.42   55.00  2.00  0.00 12.00 ggggtcaaggggaataacac
   PRODUCT SIZE: 159, PAIR ANY COMPL: 5.00, PAIR 3' COMPL: 1.00

 4 LEFT PRIMER          18   21   60.43   47.62  2.00  0.00 12.00 tttcatctggttcccttcctc
   RIGHT PRIMER        175   20   60.42   55.00  2.00  0.00 12.00 ggggtcaaggggaataacac
   PRODUCT SIZE: 158, PAIR ANY COMPL: 5.00, PAIR 3' COMPL: 2.00

Statistics
         con   too    in    in           no    tm    tm  high  high  high        high
         sid  many   tar  excl   bad     GC   too   too   any    3'   lib  poly   end
       ered    Ns   get   reg   GC% clamp   low  high compl compl   sim     X  stab    ok
Left    149     0     0     0     0     0    82    17     0     0     5     0     4    41
Right   405     0     0     0     0     0   150   117     0     3     0     0     3   132
Pair Stats:
considered 148, high end compl 9, ok 139
```



图 20.3　引物找到后 Primer3 界面

每一个输出选项的标记是有关选择的目的及 Primer3 如何利用它的文件连接。例如，敲击 Max End Stability 便可选其中一条到随后文件：

> Max End Stability
>
> The maximum stability for the five 3' bases of a left or right primer. Bigger numbers mean more stable 3' ends. The value is the maximum delta G for duplex disruption for the five 3' bases as calculated using the nearest neighbor parameters published in Breslauer, Frank, Bloeker and Marky, Proc. Natl. Acad. Sci. USA, vol 83, pp 3746~3750. Rychlik recommends a maximum value of 9 (Wojciech Rychlik, "Selection of Primers for Polymerase Chain Reaction" in BA White, Ed., "Methods in Molecular Biology, Vol. 15: PCR Protocols: Current Methods and Applications", 1993, pp 31~40, Humana Press, Totowa NJ).

在很多情况下 Max Mispriming 和 Pair Max Mispriming 输入区是很重要的，因为源序列可能含有占人类基因组 35%的散布重复序列的其中一条。用户应在挑选引物前通过 Ns 重新放置这些序列或选择 Mispriming library。(严格地讲并不是所有的错误引发都是由重复序列引起的；它可能是一个人不希望的无意扩增的任何一条序列。)然而在 www.genome.wi.mit.edu 网址的界面上只提供人类和鼠及啮齿类的重复序列文库。

最长引物的长度是受限制的，因为最接近熔链温度的模型与现实中相对短的序列是非常一致的[16, 17]。

20.2.1 Primer3 如何挑选引物

Primer3 接受许多任选项，这些任选项详细说明了哪些引物是可用的，哪些引物比其他的好。在万维网界面中用户通过文本框、复选框和下拉菜单来选择这些任选项。例如，在图 20.2 中这些包括 Mispriming Library 的下拉(在序列输入区上面)，Product Size Min、Opt 和 Max 输入区及所有 General Primer Conditions 标题下的输入区。

有些任选项具体说明哪些引物是可接受的。例如，Product Length 左边的 Min 和 Max 任选项稍微设置低些和高些就约束了产物的长度。这样的任选项就被称为 constraints(限制项)，因为它们限制了一系列可接受的引物对。还有些任选项具体说明了包括 Primer Tm Min 和 Max、Max End Stability 及 Max Mispriming 等限制项。("Tm"是熔链温度的缩写。)

其他任选项详细说明了最佳输出引物或引物对(超过那些指定的仅仅可接受的引物或引物对)的特点。这样的例子包括 Product Length Opt(最佳的)和 Primer Tm Opt 的输入。总的来说，假如 Product Length Opt 被指定，Primer3 便试图挑选一

对引物产生一个近似于指定的扩增子。对一些任选项用户不必说明最佳值，因为 Primer3 会认为它是理所应该的。例如，没有最佳 mispriming library(错误引发文库)相似性输入区，Primer3 便假定为 0。

Primer3 检查所有那些满足限制项的引物对并找出那些与最适条件最接近的引物对。Primer3 如何计算引物对符合最适条件的程度？通过设定万维网界面去相等地平衡引物长度，引物熔链温度和产物长度。(为了与仅设定引物长度和引物熔链温度的较早版本 primer3_core 有可比性。)

然而为了适应引物挑选应用的多样性，Primer3 在 formula 上是灵活的，它用这种 formula 去计算引物或引物对符合最佳条件的程度。这种 formula 的专业技术术语是 objective function。这样，你假定两个引物之间熔链温度的差异远比它们的长度、熔链温度和产物大小重要，那么你就可以使用输入页中 Objective Function Penalty Weights…部分(如图 20.4 中部分所示)告诉 Primer3 在计算最优性时考虑这些因素。在图 20.5 中标记了 Product Size Lt 和 Gt、Tm Difference 和 Primer Penalty Weight 的值来完成这种效果。(Primer Penalty Weight 是一个对 objective function 有帮助的所有 per-oligo 的调节因素。详细资料参阅在线文件。)

图 20.4 允许对 objective function 修改的部分 Primer3 万维网界面

在这个部分“Product Size Lt”和“Gt”、“Tm Difference”及“Primer Penalty Weight”已经对设定值进行了修改

图 20.5　Primer3 的“Primer Tm Min”和“Primer Tm Max”的界限完全自由的界面

20.2.2　当引物找到后对输出信息的说明

请参照图 20.3。输出的上端陈列了序列 id(如 Sequence1)和大量信息注解。输出的下一部分列出了最好的左边和右边引物，及它们的特征(起始位置、长度、熔链温度等)。接着输出(信息)还列出了输入序列和被选引物的确切资料。

接下来的信息是源序列中左引物(>>>>>…)和右引物(<<<<<…)定位的 quasi-graphical(准图形)描述和源序列的许多重要特征，在这个例子中只描述了靶序列的位置(由星号*****…标记)。序列的后面是一些额外引物对的信息。(用户可以通过在 Number to Return 输入区中输入不同的值控制输出的数量。)

最后输出(部分)还包括了题为 Statistics 的部分，这部分我们在后面会详细讨论。

20.2.3 假如没有可接受的引物怎么办?

回顾 20.1.3 节，有些情况下人们为相同的条件寻找一个好的引物而宁愿丢弃一些源序列也不愿去处理可疑的引物。大多数 Primer3 的默认选项的值都被调整来符合这些状况，限制项是严格的。考虑到严格的限制项，objective function 的细节并不重要，因为任何可接受的引物或引物对都有好的一面，但有利用潜力的引物可能会被作为不可接受的一个而丢弃。

然而，在另一种情况下，不管成本多大人们必须要设计一所给序列的引物。假如你面对的是这样一种情况，而 Primer3 又不能找到符合预置限制项的可接受的引物。这时 Primer3 就会返回到如图 20.6 所示的类似的屏幕。

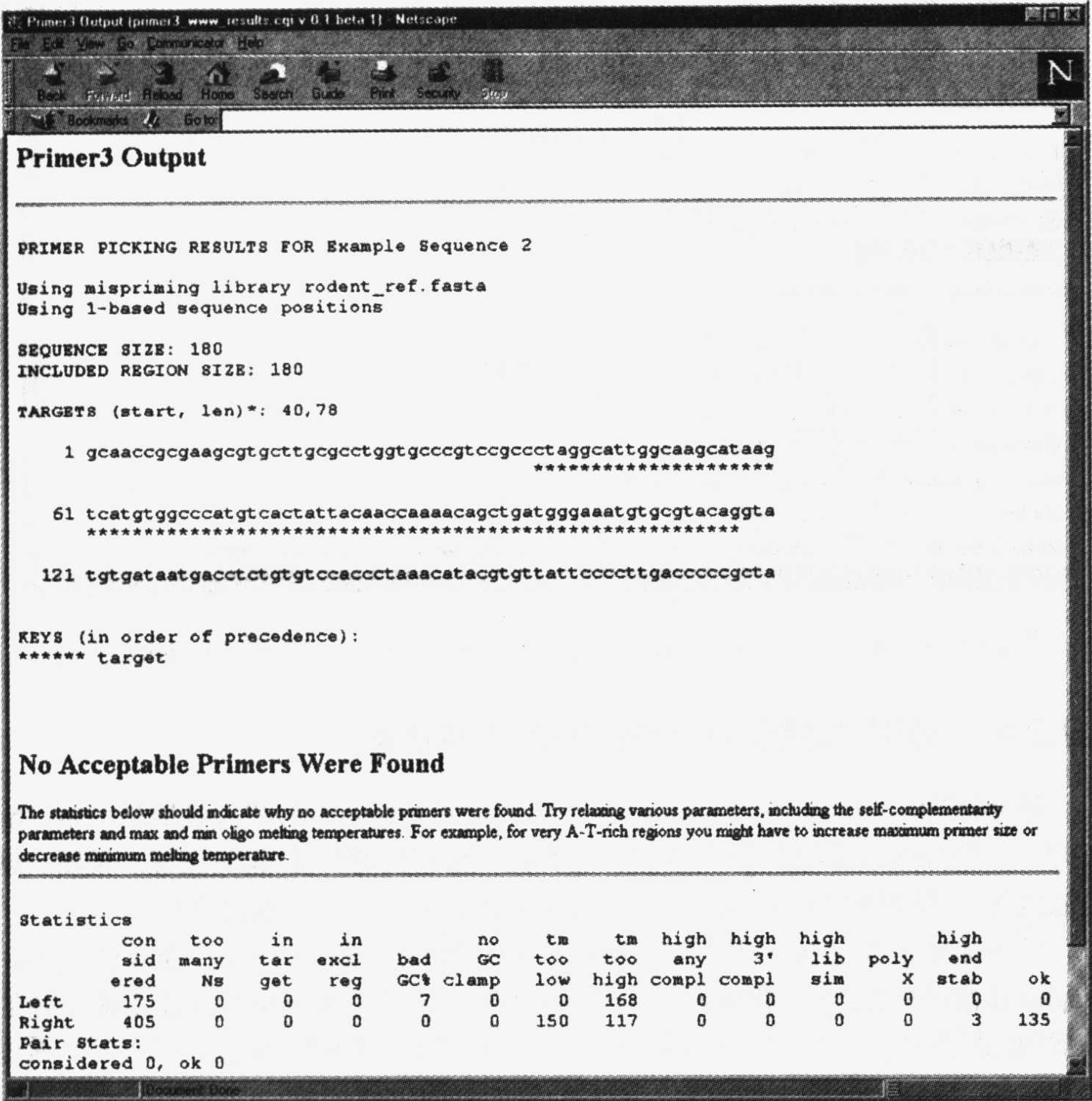


图 20.6 未找到引物时 Primer3 输出结果的界面

那么又该怎么办？这时最直觉的办法就是放宽限制项，使限制项至少是在你所处特殊情况下认为是重要的，并且限制项能尽可能防止生成不接受的引物或引物对。输出(信息)底端的 Statistics 部分表明了未被接受的引物的原因。从以上例子可清楚看出主要问题就是所有可接受引物的熔链温度过高(如题为"tm too high"栏所示)。

这里提醒一句话：Primer3 从不考虑一个因为它所处位置而为被接受的引物。因此，如果一个引物超出了涵盖的范围或由于所给序列的长度、任何特定"涵盖区"的位置及靶位和所允许产物大小的范围等原因而从未被接受，那么在 considered 栏中这个引物是不会被计算的。尽管似乎有少数引物已被考虑，你可能还想修改你的最大或最小产物大小的任选项值，或扩大涵盖区。下面就是这样一种情况的例子，它没有考虑左引物：

Statistics	con sid ered	too many Ns	in tar get	in excl reg	bad GC%	no GC clamp	tm too low	tm too high	high any compl	high 3′ compl	poly X	high end stab	ok
Left	0	0	0	0	0	0	0	0	0	0	0	0	0
Right	1401	25	0	0	9	0	884	191	0	0	0	0	292

Pair Stats：这部分表明了引物对(假设是单引物或寡核苷酸)被丢弃的原因。例如，有少数可接受的引物，当它们配对后将会产生一熔链温度过高或过低的产物。

检查 Statistics 和(一般很少)Pair Stats：这部分提示如果限制项放宽将会使引物成为被选对象。

在一些情况中(尤其当立刻放宽几个限制项时)它便希望对 objective function(目标功能)进行修改以便反映特定的引物设计目标。在图 20.6 序列中，所有可能的左引物的温度过高。继续进行的方法就是再次放松似乎已限定的限制项，如增大 Primer Tm Max 选项直到找到一个可以接受的左引物。换句话说，简单地放松所有限制项的限制可能是较简捷的方法，如图 20.5 中 Primer Tm Min 和 Max 分别设置的是 0℃ 和 100℃。用这些放宽的限制项选择的引物对是：

OLIGO	start	len	tm	gc%	any	3′	rep	seq
LEFT PRIMER	10	19	68.72	63.16	6.00	1.00	10.00	aagcgtgcttgcgcctggt
RIGHT PRIMER	175	20	60.42	55.00	2.00	0.00	12.00	ggggtcaaggggaataacac

Statistics 是：

	con sid ered	too many Ns	in tar get	in excl reg	bad GC%	no GC clamp	tm too low	tm too high	high any compl	high 3' compl	high lib sim	poly X	high end stab	ok
Left	53	0	0	0	7	0	0	0	0	9	0	0	16	21
Right	387	0	0	0	0	0	0	0	0	4	0	0	5	378

现在有可接受的引物对，但左引物和右引物之间的熔链温度有很大的差异。要缩小这种差异的一个办法就是把它作为 objective function 的一部分，如图 20.4 所示。做出这样的调整后，Primer3 便选择出如下的引物对：

OLIGO	start	len	tm	gc%	any	3'	rep	seq
LEFT PRIMER	11	18	67.83	66.67	4.00	1.00	10.00	agcgtgcttgcgcctggt
RIGHT PRIMER	180	20	66.95	60.00	2.00	2.00	10.00	tagcgggggtcaaggggaat

20.3 适合生物学程序员的 Primer3

20.3.1 安装指示

源资料可作为一个 UNIX “tar” 文件获取，该文件通过 UNIX tar 应用程序、Windows/Windows NT WinZip utility(Nico Mak Computing：`http://www.winzip.com/winzip.htm`)或通过带有 Expander Enhance utility `www.aladdinsys.com` 的 Mac DropStuff 来进行管理。要运行 primer3_core，你必须先用一个带有 POSIX 文库的编辑器编辑它并运行带有 README 解释文件的测试软件。

20.3.2 如何把 primer3_core 作为软件组成来使用的范例

在这部分中我们将举两个例子来介绍作为软件组成的 primer3_core 的使用。这些范例的代码可在 Primer3 distribution 中得到。

20.3.2.1 带有 UNIX Pipes 的 primer3_core 的使用

第一个例子是仅仅获得一个最小 perl scripting 的不太重要的应用。

这个例子说明了如何对 primer3_core 输出信息进行后处理来完成寡核苷酸设计任务。其任务就是解释说明 “overgo” (John D. McPherson, pers. comm.)，在 “overgo” 中，从 22-mer 退火的重叠区域构建一条 36-mer 的双链杂交探针并把它添到单链的尾部：

```
ACTGTGCCTGCATTTGCAGAGA
                    ||||||||
                    ACGTCTCTCCATTAATTCCATT
```

↓

```
ACTGTGCCTGCATTTGCAGAGAGGTAATTAAGGTAA
||||||||||||||||||||||||||||||||||||
TCACACGGACGTAAACGTCTCTCCATTAATTCCATT
```

我们会专门显示一个已生成引物对的编号然后设计能杂交到由引物对扩增的位置上的 overgo。下面是一条你将会用的 UNIX 命令：

```
prompt>./primer3_core < input | /.overgo.pl
```

在这条命令中 primer3_core 首先运行，从"input"文件输入，于是它的输出(信息)由 UNIX pipe(在命令条上用垂直光标"|"选定)直接发送到 perl 程序 overgo. pl。在文档编辑器中人工处理输入(信息)或(更可能地)由另一个程序生成输入(信息)。它的形式为 tag = value pairs，格式成为称为 Boulder-IO[18]：

```
PRIMER_SEQUENCE_ID=Overgo Example
PRIMER_PICK_INTERNAL_OLIGO=1
PRIMER_INTERNAL_OLIGO_MAX_MISHYB=36
PRIMER_INTERNAL_OLIGO_MIN_SIZE=36
PRIMER_INTERNAL_OLIGO_MAX_SIZE=36
PRIMER_INTERNAL_OLIGO_OPT_SIZE=36
PRIMER_INTERNAL_OLIGO_MIN_TM=10
PRIMER_INTERNAL_OLIGO_MAX_TM=90
PRIMER_INTERNAL_OLIGO_OPT_TM=70
PRIMER_INTERNAL_OLIGO_SELF_ANY=30
PRIMER_INTERNAL_OLIGO_SELF_END=30
PRIMER_INTERNAL_OLIGO_MISHYB_LIBRARY=humrep
PRIMER_PRODUCT_SIZE_RANGE=70-1000
PRIMER_EXPLAIN_FLAG=1
PRIMER_PAIR_WT_IO_QUALITY=1
PRIMER_PAIR_WT_PR_QUALITY=0
PRIMER_IO_WT_REP_SIM=1
PRIMER_IO_WT_TM_GT=0
PRIMER_IO_WT_TM_LT=0
PRIMER_IO_WT_SIZE_GT=0
```

```
PRIMER_IO_WT_SIZE_LT=0
PRIMER_NUM_RETURN=1
PRIMER_LEFT_INPUT=GAAATGTGTCCTTCCCCAGA
PRIMER_RIGHT_INPUT=GAGTTCACCCATACGACCTCA
SEQUENCE=GGATCACAACGTTTTTTGACACACCCTATAATGATGTATT...
=
```

Boulder-IO 是一种在程序之间移动 semistructured 数据的格式。Primer3 接受它的输入(信息)并(通过设定)以简单的 Boulder-IO 子集形式产生输出(信息)。Primer3 中的 README 解释文件描述了输入(信息)以及输出(信息)中所有这些 tag = value pairs 的意义。上面的输入(信息)经 primer3_core 得到的输出(信息)是:

```
PRIMER_SEQUENCE_ID=Overgo Example
PRIMER_PICK_INTERNAL_OLIGO=1
PRIMER_INTERNAL_OLIGO_MAX_MISHYB=36
PRIMER_INTERNAL_OLIGO_MIN_SIZE=36
PRIMER_INTERNAL_OLIGO_MAX_SIZE=36
PRIMER_INTERNAL_OLIGO_OPT_SIZE=36
PRIMER_INTERNAL_OLIGO_MIN_TM=10
PRIMER_INTERNAL_OLIGO_MAX_TM=90
PRIMER_INTERNAL_OLIGO_OPT_TM=70
PRIMER_INTERNAL_OLIGO_SELF_ANY=30
PRIMER_INTERNAL_OLIGO_SELF_END=30
PRIMER_INTERNAL_OLIGO_MISHYB_LIBRARY=humrep
PRIMER_PRODUCT_SIZE_RANGE=70-1000
PRIMER_EXPLAIN_FLAG=1
PRIMER_PAIR_WT_IO_QUALITY=1
PRIMER_PAIR_WT_PR_QUALITY=0
PRIMER_IO_WT_REP_SIM=1
PRIMER_IO_WT_TM_GT=0
PRIMER_IO_WT_TM_LT=0
PRIMER_IO_WT_SIZE_GT=0
PRIMER_IO_WT_SIZE_LT=0
PRIMER_NUM_RETURN=1
PRIMER_LEFT_INPUT=GAAATGTGTCCTTCCCCAGA
PRIMER_RIGHT_INPUT=GAGTTCACCCATACGACCTCA
SEQUENCE=GGATCACAACGTTTTTTGACACACCCTATAATGATGTATT...
PRIMER_LEFT_EXPLAIN=considered 1,ok 1
```

```
PRIMER_RIGHT_EXPLAIN=considered 1,ok 1
PRiMER_INTERNAL_OLIGO_EXPLAIN=considered 224,
long poly-x seq 12,ok 212
PRIMER_PAIR_EXPLAIN=considered 1,ok 1
PRIMER_PAIR_QUALITY=15.0000
PRIMER_LEFT_SEQUENCE=GAAATGTGTCCTTCCCCAGA
PRIMER_RIGHT_SEQUENCE=GAGTTCACCCATACGACCTCA
PRIMER_INTERNAL_OLIGO_SEQUENCE=ACTGTGCCTGCATTTGCA...
PRIMER_LEFT=99,20
PRIMER_RIGHT=198,21
PRIMER_INTERNAL_OLIGO=140,36
PRIMER_LEFT_TM=59.903
PRIMER_RIGHT_TM=59.981
PRIMER_INTERNAL_OLIGO_TM=72.885
PRIMER_LEFT_SELF_ANY=3.00
PRIMER_RIGHT_SELF_ANY=4.00
PRIMER_INTERNAL_OLIGO_SELF_ANY=8.00
PRIMER_LEFT_SELF_END=0.00
PRIMER_RIGHT_SELF_END=1.00
PRIMER_INTERNAL_OLIGO_SELF_END=3.00
PRIMER_INTERNAL_OLIGO_MISHYB_SCORE=15.00,MLT1b
(MLT1b subfamily)-consensus sequence
PRIMER_LEFT_END_STABILITY=8.2000
PRIMER_RIGHT_END_STABILITY=8.2000
PRIMER_PAIR_COMPL_ANY=4.00
PRIMER_PAIR_COMPL_END=1.00
PRIMER_PRODUCT_SIZE=100
=
```

overgo. pl 次级程序运行从输出(信息)而来的36-mer的杂交探针序列并产生构成overgo的22-mer的重叠片段:

```
#!/usr/local/bin/perl5 -w
$/="\n=\n";  # Set the record terminator.
while (<>)  {
  %rec=split /[=\n]/;     # A DANGEROUS approach
                          # to parsing the sequence.
  for (keys %rec)  {$rec{$_}=~s/\n//}
```

```
  $seq=$rec{'PRIMER_INTERNAL_OLIGO_SEQUENCE'};
  next unless $seq;
  print "MARKER\t\t$rec{'PRIMER_SEQUENCE_ID'}\n";
  $left  =substr($seq,0,22);        # Get left oligo.
  $r     =substr($seq,14);          # Get the right
                                    # oligo,
  $right=reverse($r);               # reverse it,and
  $right=~tr/GATC/CTAG/;            # complement it.
  print "LEFT_MID_OLIGO\t$left\n";
  print "RIGHT_MID_OLIGO\t$right\n";
  print "MAX_SCORE\t
  $rec{'PRIMER_INTERNAL_OLIGO_MISHYB_SCORE'}\n";
  $gc   =($seq=~ tr/GC/GC/);
  printf "GC_content\t%d%%\n\n",$gc * 100 / 36;
}
```

一条$/="\n=\n"的程序语句告诉 perl 每一条记录是由自身程序线上一个"="符号来终止(这个符号是 Boulder-IO 的标准终止符)。`%rec=split/[=\n]/;` 程序语句将 Boulder-IO 记录解释到 perl hash `%rec` 中。这种解释分析输出(信息)的方法要求我们知道"="不会出现在任何 Boulder-IO tag = value pair 的 value 部分中。为获得更稳健的状态，可使用 Lincoln Stein's perl Boulder 模块(在 http://www.genome.wi.mit.edu/genome_software/other/boulder.html 网址上可得到)。用这种模块 overgo. pl 会被改写如下：

```
#!/usr/local/bin/perl5 -w
use Boulder :: Stream;
$in=new Boulder :: Stream;
while ($rec=$in->read_record()) {
  $seq
     = $rec->get('PRIMER_INTERNAL_OLIGO_SEQUENCE');
  next unless $seq;
  print"MARKER\t\t",
  $rec->get('PRIMER_SEQUENCE_ID') , "\n";
  $left = substr($seq,0,22);   # Get the left
                               # oligo.
  $r    = substr($seq, 14);    # Get the right
                               # oligo,
  $right= reverse($r);         # reverse it,and
```

```
    $right = ~tr/GATC/CTAG/;    # complement it.
    print "LEFT_MID_OLIGO\t$left\n";
    print "RIGHT_MID_OLIGO\t$right\n";
    print "MAX_SCORE\t",
    $rec->get('PRIMER_INTERNAL_OLIGO_MISHYB_SCORE'),
    "\ n"  ;
    $gc    =($seq=~tr/GC/GC/);
    printf "GC_content\t%d%%\n\n",$gc * 100/36;
}
```

用 Boulder 模块是可取的，因为它比较稳健。即使有人输入“=”，例如是 PRIMER_SEQUENCE_ID 的值，它也能正确的运行。唯一不便的就是在你使用它前必须得到 Boulder 模块。以上输入(信息)运行后的输出(信息)是：

```
MARKER                  Overgo Example
LEFT_MID_OLIGO          ACTGTGCCTGCATTTGCAGAGA
RIGHT_MID_OLIGO         TTACCTTAATTACCTCTCTGCA
MAX_SCORE  15.00,       MLT1b(MLT1b subfamily)...
  consensus sequence
```

20.3.2.2　从 perl 中调取 primer3_core

第二个例子是 Primer3 万维网界面的 primer3_core。这个代码片段是从 CGI script 作为该界面的补充成分的 primer3_www_results. cgi 适配来的。CGI 模块可从 http://www.genome.wi.mit.edu./ftp/distribution/software/WWW/得到，Primer3_www. cgi 称作 primer3_core，这部分带有一个要求格式化输出(-format_output)的标志，然后选中 primer3_core 输出(信息)：

```
#!/usr/local/bin/perl5 -w
...
use FileHandle; # Standard part of perl distribution
use IPC::Open3; # Standard part of perl distribution
use CGI;
    ...
    $query = new CGI;
                    # $query now contains the parameters
                    # to the cgi script
    ...
    my @names = $query->param;
    ...
```

```
        for(@names)  {
                next if...# Some cgi parameters do not get
                          # sent to primer3_core
               ...
                $1ine = "$_= $v\n";
                push @input,$1ine;#Save a Boulder-IO line for
                                  #primer3_core's  eventual
                                   consumption.
    }
    my $cmd = "./primer3_core -format_output -strict_tags"
    my $primer3_pid;
    my($childin,$childout)=(FileHandle->new,FileHandle->
       new);
    {
       local $^W = 0;
       $primer3_pid = open3($childin, $childout, $childout,
        $cmd);
    }
if(!$primer3_Pid)  {
    print "Cannot excecure $cmd:<br>$!\n$wrapup\n";
    exit;
    }
    print "<pre>\n";
    print $childin @input;
    $childin->close;
    my $cline;
    while($cline = $childout->getline){
      if($cline = ~ /(.*)start  len   tm  gc%  any  3\'seq/)
      {
        # Grap a particular line and
        # add hyperlinks to it:
        $cline = $1
            ."<a href = \"$DOC_URL#PRIMER_START\">start</a>"
            ."<a href = \"$DOC_URL#PRIMER_LEN\">len</a>"
            ."<a href = \"$DOC_URL#PRIMER_TM\">tm</a>"
```

```
        ."gc%   any    3\' seq\n"
    }
    print $cline;
  }
  print "</pre>\n";
  waitpid $primer3_pid,0;
  if($? ! = 0 && $? ! = 64512)  { # 64512 == -4
    ... # primer3_core exited with
        # an error code;alert the browser.
  }
```

当然在$cmd 中格式化的输出信息标志(formated_output flag)在这个范例中并不是工作中的必须示范部分。Script 已经解释(parse)了 Boulder-IO 输出信息并接着以其他方式格式或运行这些信息。

20.3.2.3 primer3_core 作为软件组成的其他应用

以上两个范例说明了在不太重要的 Unix pipeline(流水线)(overgo 设计范例)中如何把 primer3_core 作为一软件组成来使用，以及如何使用 perl's open3 命令来启动执行 primer3_core 程序并选取其输出(信息)进行进一步处理。一个对程序来说比使用 open3 更简单的适中方法就是简单地使用 perl open 命令然后返回到未修改的 primer3_core 输出，如：

```
  if(!open(PRIMER, "|$cmd")) {
    print "Cannot execute <pre>$cmd\n</pre>\n$wrapup";
    return;
  }
  print PRIMER @input;
  close PRIMER;    # primer3_core's output is the same
                   # as this script's output.
  if($? !=0 &&$?!= 64512){#64512== -4
     ... # $cmd exited with an error code.
  }
```

在 Whitehead institute 中我们已经使用 primer3_core，它作为 industrial-strength 引物设计 pipeline(流水线)包括向量剪裁(鉴定和向量臂的电子拆除)、微卫星重复鉴定和向量污染物的自动筛选。在 pipeline 中我们使用它并添加恒定的 5′尾端到每一条引物及在 pipeline 中寻找一条序列上一系列的扩增子。为了最后这一条应用，我们在 primer3_core 输入中设定 PRIMER_FILE_FLAG = 1，它能指导 primer3_core 产生包含所有可接受的左、右引物的文件。那么不同程序就可从这

些所列引物中挑选引物去生成 tiling。

20.3.3 效益考虑

Primer3 的运行时间对万维网界面的用户是很少用的术语。然而大容量应用 primer3_core 的用户应该知道约定运行时间这个因素。选择单个引物最昂贵的操作是对 mispriming 或 mishyb 文库(每个寡核苷酸所需的实际时间是与文库大小呈线性关系)检查。如果 Primer3 检查大数量的引物对，为自身补充检查寡核苷酸对，那接下来最昂贵的操作是对寡核苷酸自身补充的检查。

Primer3 的运行时间还依赖于要选择引物序列的大小。在 10kb 大小序列任何地方挑选单个引物对大约花费 10 倍于在 1kb 大小序列任何地方挑选单个引物的时间(其他选择项都是一样的)。

接下来是 Primer3 的运行时间的其他决定因素：

(1) 放松寡核苷酸的限制项。在检查较 expensive-to-compute 特征(如与 mispriming 文库输入相似)前 Primer3 排除基于便宜计算的引物(如寡核苷酸的熔链温度)以便放松 cheap-to-compute 限制项限定的评定。

(2) 引物可接受的位置(还要考虑产物大小的限制项)。这一项与前一条相似。Primer3 不执行昂贵的操作去对由于它们的位置而从未作为可接受引物对一部分的引物进行特征描述。

(3) PRIMER_FILE 输入标记。这一标志使得 Primer3 对每一个特征计算，包括 mispriming 相似性和每一个可接受引物的自身补充。

(4) 计算 objective function 的成本。有两个次要因素。

a. Objective function 依赖于寡核苷酸或引物的 expensive-to-compute 的特征，例如，一对引物之间 mispriming 或 mishyb 文库或补充的相似性。在这个例子中 Primer3 必须对所有可接受引物进行昂贵的计算。

b. Objective function 依赖于引物对 per se 的特征，如产物熔链温度或产物大小。在这个例子中 Primer3 必须计算是否每一个引物是可接受的，通常哪一个需要通过昂贵的计算来决定可接受性。

(当 Objective function 既不依赖于每个引物昂贵的特征也不依赖于引物对特征时，Primer3 便对其查找进行组织便于它对最好引物的昂贵的限制项进行核对。)

致谢

Primer3 和 Primer3 万维网界面的开发由 Howard Hughes Medical Institute、National Institute of Health 和 National Human Genome Research Institute 提供项目资助，项目授权号为 R01-HG00257(David C. Page)和 P50-HG00098(Eric S. Lander)。

我们诚挚地感谢 Digital Equipment Corporation 的支持，该公司提供给我们用于 Primer3 大量开发的 Alphas，衷心感谢 Centerline Software 公司，该公司的 TestCenter memory error、memory leak 和 test coverage checker 帮助我们发现并纠正 Primer3 中别的方面大量的潜在错误。

Primer3 是 Whitehead Institute 最近完成的大量引物挑选程序 ，该程序用 Primer0.5 启动[13]。Primer3 是作为软件组成的 Primer 0.5 的再版启用；Primer3 的设计很大程度上依赖于 Primer0.5、Primer v2(Richard Resnick)的设计和 Richard Resnick 为 Primer v2 设计的万维网界面的计划。

感谢 Alex Bortvin，Mark Daly，Nathan Siemers 和 William J. Van Etten 为本章修改手稿。

(蒋红霞 译)

参 考 文 献

[1] Dieffenbach, C. W. and Dveksler, G. S. (1995) *PCR Primer A Laboratory Manual.* Cold Spring Harbor Laboratory Press, Cold spring Harbor, NY.

[2] Innis, M. A., Gelfand, D. H., Sninsky, J. J., and White, T. J., eds. (1990) *PCR Protocols A Guide to Methods and Applications.* Academic Press, San Diego, CA.

[3] Rychlik, W. (1993) Selection of primers for polymerase chain reaction, in *Methods in Molecular Biology, vol. 15: PCR Protocols: Current Methods and Applications* (White, B. A., ed.) Humana, Totowa, NJ, pp. 31-40.

[4] Wetmur, J. G. (1991) DNA probes: applications of the principles of nucleic acid hybridization. *Crit. Rev. Biochem. Mol. Biol.* **26,** 227-259.

[5] Schuler, G. D. et al. (1996) A gene map of the human genome. *Science* **274,** 540-546.

[6] Wang, D. G. et al. (1998) Large-scale identification, mapping and genotyping of single-nucleotide polymorphisms in human genome. *Science* **280,** 1077-1082.

[7] Harbison, S. and Steele, G. (1995) *C A Reference Manual,* 4th ed. Prentice Hall, Englewood Cliffs, NJ.

[8] Dougherty, D. (1991) *POSIX Programmer's Guide.* O'Reilly, Cambridge, MA.

[9] Gundavaram, S. (1997) *CGI Programming with Perl.* O'Reilly, Cambridge MA.

[10] Stein, L. D. (1997) *How to Set Up and Maintain a Web Site,* 2nd ed. AddisonWesley, Reading, MA.

[11] Wall, L., Christiansen, T., and Schwartz, R. L. (1996) *Programming Perl,* 2nd ed. O'Reilly, Cambridge, MA.

[12] Rychlik, W. and Rhoads, R. E. (1989) A computer program for choosing optimal oligonucleotides for filter hybridization, sequencing and in vitro amplification of DNA. *Nucleic Acids Res.* **17,** 8543-8551.

[13] Daly, M. J., Lincoln S. E., and Lander E. S. (1991). "*PRIMER*", unpublished software, Whitehead Institute/MIT Center for Genome Research. Available at `http://www.genome.wi.mit.edu/ftp/pub/software/primer.` 0.5, and via anonymous ftp to `genome.wi.mit.edu, directory/pub/software/primer.0.5.`

[14] Hillier, L. and Green, P. (1991) OSP: an oligonucleotide selection program. *PCR Meth. Appl.* **1,** 124-128. Documentation available at `http://genome.wustl.edu/gsc/manual/protocols/ospdocs.html.` OSP is available from the author on request.

[15] Smit, A. F. A. (1996) The origin of interspersed repeats in the human genome. *Curr. Opin. Genet. Devel.* **6,** 743-748.

[16] Breslauer, K. J., Frank, R., Bloeker, H., and Marky L. A. (1986) Predicting DNA duplex stability from the base sequence. *Proc. Natl. Acad. Sci. USA* **83,** 3746-3750.

[17] Rychlik, W., Spencer, W. J., and Rhoads, R. E. (1990) Optimization of the annealing temperature for DNA amplification in vitro. *Nucleic Acids Res.* **18,** 6409-6412.

[18] Stein, L. (1997) How perl saved the human genome project. *Dr Dobb's Journal* Spring 1997 Special Report on Software Careers. Available at `http://www.ddj.com/ddj/1997/careers1/stei.htm`.

21 利用万维网装备分子生物学实验室

MaryAnn Labant　Roger Anderson

21.1 引言

因特网的世界范围性使购买实验室研究设备的电子市场得到了发展。万维网的发展为供应商提供了有效的工具，使他们能不断地提供有关产品、价格、功能以及订货情况的最新信息。信用卡的使用(一般指 P 卡)是开启贸易新时代的必需工具之一。

“网上的贸易宗旨是提高效率、降低成本、使贸易进一步全球化”(载自 Cyber Commerce——internet Tsunami, Goldman Sachs technology Report, Aug.4, 1997)。今天许多机构都在逐步改进其处理信息的方法，尤其在金融系统方面。同时供应商也正在逐步改善交流信息的方式，尤其是电子传媒的使用。买卖双方工作方法的系统性改进直接影响了实验室研究人员，因为他们是这个过程的最终顾客。反过来实验室研究人员也直接影响着这些新方法新系统能否成功执行。顾客竞争和变化正在迫使供应公司提供更好的合作和支持，而这些的最终受益者将是研究人员。

新贸易时代正在不断发展，机构和研究人员在购买产品和服务时仍有很多选择。本章的目的是针对一些分子生物学实验室研究人员当前所面对的一些选择和限制提出一些见解。

21.2 网络上变化不断

在世界变化迅速的今天(企业合并、地区代码及地址的变化等)网上过时信息的存在是不可控制的。许多公司和个人建立的信息性网站并没有持续运作。这些网站在网上或存在或已消失，结果你登录时，不是无法链接就是发现过时的信息。在一次对 25 个信息网站测试中发现超过 35%的网站有半年未更新信息。由于公司资金不足或管理网站的人员由于毕业、职位调动或是失去兴趣等原因，许多废弃的网站通常没被作者清除。所以在你浏览一个网站之前应该检查一下更新记录，以确保你查到的是最新信息。

网上发布信息的一个优点是内容生动，缺点是随着网络技术人员开发新的工具，网站的设计、背景、内容的变化频繁发生，有许多网站开发工具和更多开发

技术人员，这种结合技术使网站的设计和实际运用方面有很大的差异。没有两个供应商、团体、机构、公司或杂志有相似的网站。菜单、链接方式、布局内容都是不同的。由于顾客口味和需求的变化，为了与顾客之间架起沟通的桥梁，了解最新需求，公司也改变他们的网站。

21.3 在网上查询产品信息

今天网上有充足的信息可以利用，并没有规定你在一个网站必须发现什么，所以你只需登录随便看看，如果你发现了自己喜欢的网站可以记下来，这样下次访问时就更容易了。供应商开始建立网站时，用固定网页，而现在利用更复杂的方法和数据来制作能随时提供新信息的网页来满足顾客的需求。举个例子，在写这篇文章时，Millipore 公司作为过滤产品市场的领导者，它的网站提供了深入彻底的搜索技术信息、附有多种语言的目录，以及在线订货服务。

公司正在越来越多地利用搜索引擎，这给他们的网站增色不少。这些搜索程序有一些相同的特征。如引文中的关键词放在一起搜索。分开写搜索时，不会有引文被搜到，当放到一起写搜索时，会发现很多引文。人们正在试图推出更好的搜索程序。当然如果你所登录的网站没有搜索格式中所需的大量有用信息，这一切都是无用的。许多网站的搜索程序说明你都应该看一下，这会使你的搜索容易一些。

在线产品的范围或是十分广泛，或是十分有限。一些供应商只列出他们的最新产品或最畅销产品，他们可能对提供全部产品的信息的网站无兴趣，也可能他们的产品信息在电子格式下不易更新。因而，即使你知道某公司出售某产品，你也未必能搜索到结果。公众压力可能会改变这种状况，因为使用者从可供选择的市场来购买产品，他们对产品的信息需求对供应商来说非常重要。反馈意见会被认真对待，尤其是有正当理由且以专业的方式提出意见，让他们知道你的感受。

供应商的网站复杂程度和能力也大不相同，大多数供应商的网站有提供需求信息和资料的能力。一些供应公司允许使用信用卡在线订购产品，而其他公司则进行了限制：只有具有账户的顾客才具备在线订购的权利。直接销售的供应商正在谋求与出版商联络实现网上销售。

以前，供应商在网站有“有用的和感兴趣的链接”这一部分，而现在这些链接大部分已不存在了。作为替代，他们继续维持和扩展起独有的在线技术支持。

除了供应商的网站之外，还有途径可获得技术产品的信息。贸易出版物和协会的网站就是个很好的例子(表 21.1)。而一些网站，如“科学”杂志的网站只在广告中列出信息。其他如“生物医学产品新闻出版物”的网站和国际科学交流出版公司的网站，有在线购物指南可随时参考，购物指南对供应商的具体产品区域

的服务验证十分有帮助。

表 21.1　杂志和相关站点的一些例子

出版物杂志或相关者	目录名	站点网址
Biomedical Products News	Life Science Lab Reference Supplier Directory	www.biomedprod.com
International Scientific Communications	Lab Crawler-ISC Buyer's Guides	www.iscpubs.com
The Scientist	Lab Consumer	www.the-scientist.com
American Chemical Society	Lab Guide	pubs.acs.org
(美国化学会)	General Information and links	www.acs.org
	Products	www.cas.org
	Product Information Network(PIN)	www.chemcenter.org
Bio Techniques	Buyer's Guide and BioMall	www.biotechniques.com
Nature	Buyer's Guide and Biotechnology Directory	www.nature.com
Science	Electronic Marketplace	www.sciencemag.org

本表是一个通用例子，并未提供出版物和相关站点的综合列表

除了贸易出版物和协会的网站之外，许多其他独立的网站已经发展成为把科学的市场的产品和供应商信息联系到一些网站上，这些网站从提供产品信息的数量、所提供的类型、分类方法搜索机制到交流信息的能力，或者直接订货的能力都是不同的。当寻找最初的产品信息或供应商时，这些网站是最有用的。20 世纪 90 年代中期，许多独立发展的网站直接和其他网站链接(表 21.2)，这是非常重要的，因为高效的搜索引擎还没应用。这些最初的站点仍然存在，但在几年之内没被升级。Pedro's Tools 是其中较好的一个，一直延续这一标题，最后的一次升级时间是在 1996 年 6 月。

表 21.2　能用于查找供货商及商品信息的独立站点的一些例子

网址	描述
www.atcg.com	3000 家供应商提供的生物学和分析研究所用的 600 000 种生命科学、MRO 和办公详细产品信息列表
www.biosupplynet.com	2700 家供应商的 15 000 种产品信息列表
www.bio.com	化学和生物学研究应用软件列表
www.chemconnect.com	化学试剂、供应商和文献资料列表
www.biolinks.com	科学供应商和其他有用信息列表
www.sciquest.com	超过 300 000 种产品和临床工业产品信息列表
www.bio.net	BioSci 是一种电子交流平台。其中的方法和试剂讨论组常常涉及新的产品
www.antibodyresource.com	抗体供应商、数据库和其他资源列表
www.cato.com	生物技术和制药工业产品和服务目录列表

这些站点信息来源于写作时的出版文献

一个关于独立站点的警告性提示是关于信息有效性的决策。这并不要求揭示如发起人与应用者或其他受益群体的关系。这可能成为一种误导，任何因特网上找到的信息都不应该当作完全正确而拿来使用。

在网络资源的不断增长中，我们鼓励你能成为一名积极的参与者。

21.4 发现了所需要的产品，你能从网上购买吗？

你最终找到了你曾上网查找过的产品，并且这个站点允许在线订购。既然你找到了你想要的产品，你能通过在线购买吗？从站点购物的能力不仅取决于你所在单位购买规定，而且取决于不同公司所能接受的不同的付款方式。

21.4.1 组织的购买程序

在科研环境下，研究人员具有广泛的购物自主权。然而，他们也需要遵守相关的协定，组织的规定以及为他们所买的产品负责。经常是一个单位指派某个人负责购买物资，让该人设计他们自己的进货渠道。但不幸的是，这些人可能比在进货部花费双倍的努力，却不免为他们的物资花高价。

虽然研究人员可能会反对，但在任何一个单位组织内，无论是供应商的网站、个人站点或是一个单位组织范围内的采购系统，监控都是非常挑剔的。管理者必需监督花费、洽谈以及与合同相关的账目，核实与政府规定的服从状况，并监督供应商的执行情况。

由于这些保护，供应商必须能证实订货的人是被授权购货的人。供应商也需要得到运输费用的付款保证。买卖双方都需要并且要求安全与隐私的保护。单位组织内的电子目录和商业系统，为用户提供灵活性、隐私性的保证，以及敞开的联系，允许单位组织保持监控并能适时调整以适应购物需求的变动，并且同时减少欺诈的发生。

有其他因素影响在哪里购买和能否购买商品。2000 年的计算机杂志对单位的做法提出了挑战，财务申请正在被修改或取代。使事情更复杂的是，许多组织决定，一项主要的基础设施申请升级保证了重新启动的努力与过程的重新设计。这种在你单位内全新的购物方法的实施，依赖于最后的全部过程的分析什么时候能完成。这将影响每个人，包括研究人员、管理部门以及供应商。

把更多的产品选择责任以监控订货交给研究人员或其他终端用户的手中成为一种趋势。各种单位用来完成分散购买权的方法有：运用电子目录、使用供应商网络、电子商务或以网络为基础的订购系统，以及进货卡。所有这些系统应与财政申请共同作用，以在订货被发送给供应商去履行前，保证足够的资金到位。虽然听起来很繁琐，官僚主义和耗费时间，但在事实上，今天信息流通更加具有有效性、影响力以及顾客至上的特点。

如果一个系统设计正确，订货的人不需要了解背后繁琐的数据传送过程。但他们必须认识到好处，其中最主要的一条是正确估计账目平衡。在财政季节或年度结束时，每个人都节约时间，因为资金平衡继续保持，消除了“或使用或放弃”的购货原则，避免了透支的激烈争论。

21.4.2 市场对个体供应商站点

今天，在线订购有许多行之有效的选择，范围从集体市场到个人供应商站点。一些市场和供应商使用 EDI(Electronic Date Interface)作为基本目录建设区块。使用 EDI 的优点在于数据能够允许展示产品的有效性和价格而实时交换；缺点是由于数据空间的大小对产品的描绘也许不能被破解，例如，一个项目条款下有 30 个特点，以便于其他信息资源用来作为一个产品决策。

其他的营销系统是使用一个 mall scenario 或者登录个体供应商的网站，使信息容易接受，通过一个共同的通道，经常是站点上的一个画面，或是通过个体供应商的产品信息，并在 mall 里提供每个供应商一个 store。在第一种方法里，研究人员必须去了解供应商的每个站点上的内容，这与了解供应商的书面目录内容是相似的。后面的 mall scenario 相似地组织每个 store，虽然用户必须在 mall 里逐个 store 去寻找。在一个 mall model 里，给浏览器不断增多的命令，是计算机资源的一个负担。

在讨论的最后部分，我们将集中讨论在 www.atcg.com 中我们是怎样建设市场资源的。这个资源有两部分构成：

- ATCG 目录：一个稳定的信息源。
- 产品窗口™：电子商务服务。

ATCG 目录是一个综合市场。目录是一个可查的信息资源，在一个单一产品的数据库中，包含了 100 万条的来自于供应商及厂家的产品信息。我们已经努力使它成为一个完整的相关产品信息资源，使用户能快速找到并比较产品的供应，从供应商一次订货。

资源的开放建设，允许供应商和科学数据编辑者在目录中保持产品的数据(表 21.3)。

表 21.3 一些产品分类目录

• Antibodies	• Clinical Supplies	• Lab Organisms	• Office Supplies
• Apparel	• Columns	• Labware and Equipment	• Photographic Materials
• Books	• Filters and Membranes	• Libraries	• Plasticware
• Broths, Media and Sera	• Gels and Gel Materials	• MRO Supplies	• Proteins and Peptides
• Chemicals	• Glassware	• Modifying Enzymes	• Restriction Enzymes
• Chromatography Supplies	• Kits	• Nucleic Acids	• Vectors

建站的目的是在目录的每个产品列表中提供足够的相关技术信息，以便于用户能够根据在目录中所能找到的这些信息做出购买的决定。同时链接有其他的相关信息，它能帮助你顺利做出购买产品的决定，例如，缓冲剂的组成成分。每个产品列表中都包含供应商的姓名，并可链接到供应商的包括个人信息在内的信息页，并还可以链接到供应商的网址和 e-mail 地址(图 21.1)供应商可选择图表或目录并回复到他们的网址上。

ANDERSON UNICOM GROUP, INC.
Restriction Enzyme Search

Your Name (username)
Institutional Administrator Account
Your Institution's Name Here

EcoRI*

Alternate Names For This Product: EcoR I

Type	Methylation Sensitivity	Overhang	Unit Type	Cut Site
II	NS	5'	Activity	G/AATTC

Isoschizomers: None

48 available products for EcoRI* :

Supplier	Cat. #	Mfr.	Note	Buffer	Conc.(U/µl)	Use Temp.	Units	List Price	Price/Unit	
AAB	H-01-05000			A	40 U/µl	37 °C	5000.000 U	$19.00	$0.0038	Select
AAB	L-01-05000			A	10 U/µl	37 °C	5000.000 U	$19.00	$0.0038	Select
AAB	H-01-25000			A	40 U/µl	37 °C	25000.000 U	$81.00	$3.2399	Select
AAB	L-01-25000			A	10 U/µl	37 °C	25000.000 U	$81.00	$3.2399	Select
AAB	H-01-50000			A	40 U/µl	37 °C	50000.000 U	$128.00	$2.5600	Select
AAB	L-01-50000			A	10 U/µl	37 °C	50000.000 U	$128.00	$2.5600	Select
Amersham	E1040Y		ECORI	H	8-12 U/µl	37 °C	5000.000 units	$28.00	$5.5999	Select
Amersham	E1040Z		ECORI	H	8-12 U/µl	37 °C	10000.000 units	$43.00	$0.0043	Select
Amersham	E1040ZH		ECORI	H	>40 U/µl	37 °C	10000.000 units	$42.00	$4.1999	Select
Amersham	E1040XH		ECORI	H	>40 U/µl	37 °C	50000.000 units	$139.00	$2.7799	Select
Amersham	27-0854-03			NS		NA	5.000 u or 1.0 EA	$31.00	$6.2000	Select
Amersham	27-0854-04			NS		NA	25.000 u or 1.0 EA	$122.00	$4.8799	Select
Amersham	27-0854-18			H	50-100 U/µl	37 °C	25000.000 U	$116.00	$0.0046	Select
CHIMERx	2150-01				2-15 U/µl	37 °C	5000.000 U	$30.00	$6.0000	Select
CHIMERx	2150-01A			NS		NA	10000.000 U	$40.00	$4.0000	Select
CHIMERx	2150-02				2-15 U/µl	37 °C	25000.000 U	$110.00	$4.4000	Select
CHIMERx	2150-02 HC				40-60 U/µl	37 °C	25000.000 U	$110.00	$4.4000	Select
CHIMERx	2150-02A			NS		NA	50000.000 U	$158.00	$0.0031	Select
Life Tech/Gibco	15202013		Cloned	3	8-12 U/µl	37 °C	5000.000 U	$27.00	$5.4000	Select
Life Tech/Gibco	15202021		Cloned	3	8-12 U/µl	37 °C	20000.000 U	$79.00	$3.9500	Select
Life Tech/Gibco	15202039		Cloned	3	50 U/µl	37 °C	20000.000 U	$86.00	$0.0043	Select
Life Tech/Gibco	15202120		Cloned	3	8-12 U/µl	37 °C	60000.000 U (3 x 20000.0 U)	$179.00	$2.9833	Select
NEB	101CS		Cloned by NEB	U	100 U/µl	37 °C	10000.000 U	$44.00	$4.4000	Select
NEB	101S		Cloned by NEB	U	20 U/µl	37 °C	10000.000 U	$44.00	$4.4000	Select
NEB	101CL		Cloned by NEB	U	100 U/µl	37 °C	50000.000 U	$176.00	$3.5200	Select
NEB	101L		Cloned by NEB	U	20 U/µl	37 °C	50000.000 U	$176.00	$3.5200	Select
NEB	101CXL		Cloned by NEB	U	100 U/µl	37 °C	200000.000 U	$400.00	$0.002	Select
NEB	101XL		Cloned by NEB	U	20 U/µl	37 °C	200000.000 U	$400.00	$0.002	Select
Promega	R6011		B/W Cloning Qualified.	H	8-12u/µl	37°C	5000.000 U	$30.00	$6.0000	Select
Promega	R6017		B/W Cloning Qualified.	H	8-12u/µl	37°C	15000.000 U	$65.00	$4.3333	Select
Promega	R4014		B/W Cloning Qualified.	H	40-80u/µl	37°C	25000.000 U	$95.00	$0.0038	Select
Promega	R6012			H	8-12u/µl	37°C	25000.000 U (5 x 5000.0 U)	$120.00	$4.7999	Select
Promega	R6018			H	8-12u/µl	37°C	75000.000 U (5 x 15000.0 U)	$260.00	$3.4666	Select
Stratagene	500480		Cloned	U1½	10-120 U/µl	37 °C	10000.000 U	$39.00	$3.8999	Select
Stratagene	500481		Cloned	U1½	10-120 U/µl	37 °C	50000.000 U	$149.00	$0.0029	Select
Stratagene	500489		Cloned	U1½	10-120 U/µl	37 °C	50000.000 U	$149.00	$0.0029	Select
WAKO	314-00112			H	5-20 U/µl	37 °C	12000.000 U	$94.75	$7.8958	Select
WAKO	314-01751			H	50-200 U/µl	37 °C	12000.000 U	$94.75	$7.8958	Select
WAKO	310-00114			H	5-20 U/µl	37 °C	60000.000 U (5 x 12000.0 U)	$249.25	$4.1541	Select
WAKO	318-00115			H	5-20 U/µl	37 °C	60000.000 U (5 x 12000.0 U)	$427.00	$7.1166	Select

Main Menu:

Home | Product Search | Supplier Profiles | User Profile | View Cart
Links | Manual | FAQ | Contact Us | About Us | Log-Out

JAVA ON

Click the clock for help   Site images Off

Document: Done

图 21.1　用 48 种不同供应商提供的产品中对限制性内切核酸酶 *Eco*R I 的搜索

我们当前的市场并非按供应商分类，而是按产品目录分类的，创建了大量的可寻找和可比较的信息资源。所以，使用者可以将一页中许多供应商的产品进行比较，并选择他们需要的产品。所有的信息可以按价格、单位、单价或供应商分类，以便使使用者能最后作出决定。

我们可以通过以下 4 种方式寻找产品：产品种类、目录数量、正文和供应商(图 21.2)。产品目录选用一种标准的网页来帮助使用者。每个产品都制定了选择标准，每一个标准列表有多项条款可被选择，其中当然包括供应商信息网站上的产品资源。

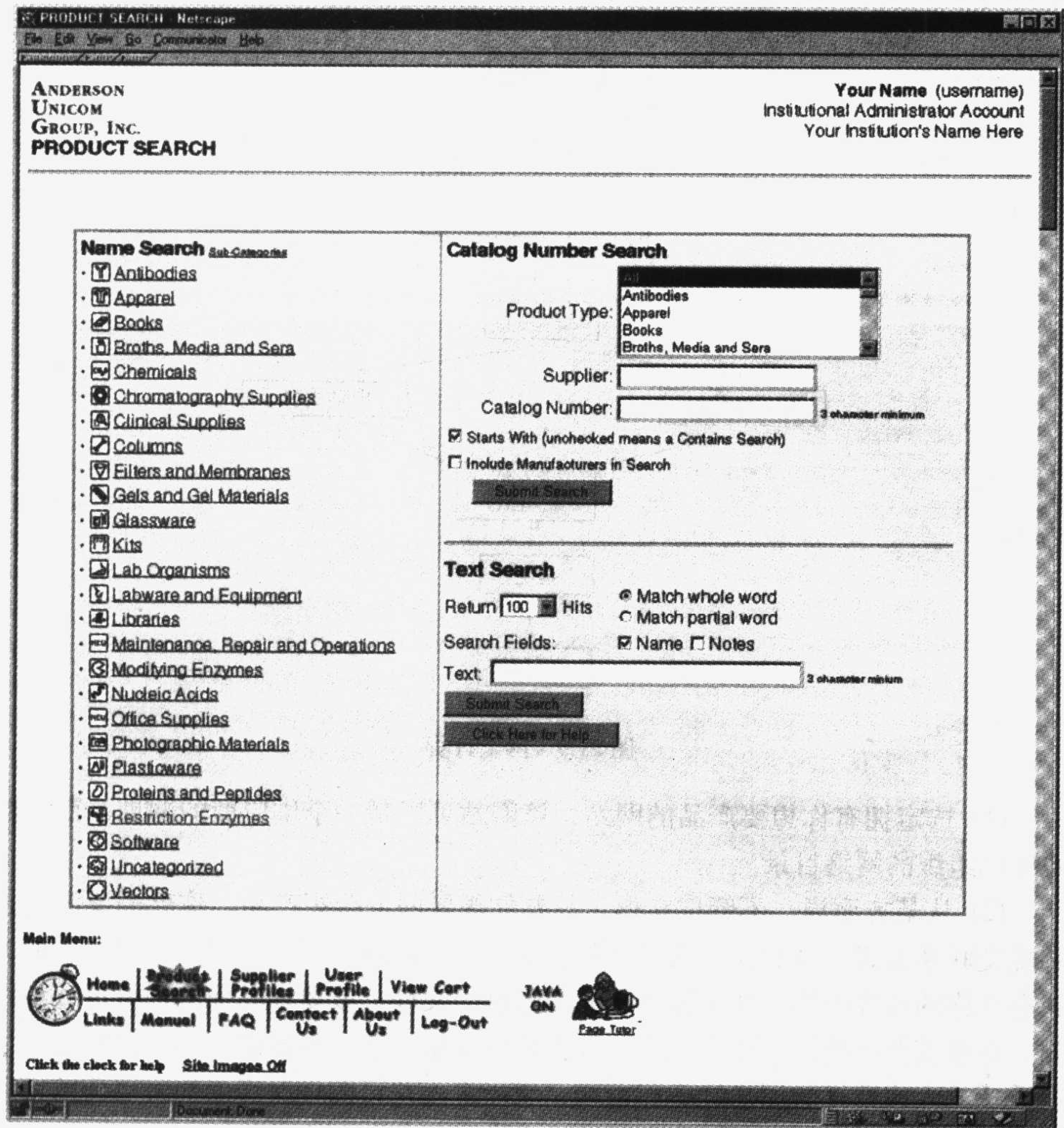


图 21.2　ATCG 类的搜索界面提供了 4 种发现产品的方法：
通过产品类别、产品分类号、供应商及文本搜索等

作为面向科学界的一个服务性资源，对进入者公开，所有进入者都可以查询产品信息。美国和加拿大的注册用户还可以看到价格列表。埃德森联合会的还为客户提供需 Product Window™ 的开发与执行的电子目录,这也是我们的商业服务。在面向顾客的组织里，有组织和执行作用的供应商也被加到市场上去，也就是与在特定组织里有权威性的使用者协商价格。

21.4.3 在线订购

正如埃德森联合会或其他可用的个人供应商的网站，网络系统可使使用者全天 24 小时订购。下面是在线订购的简要程序(图 21.3)：

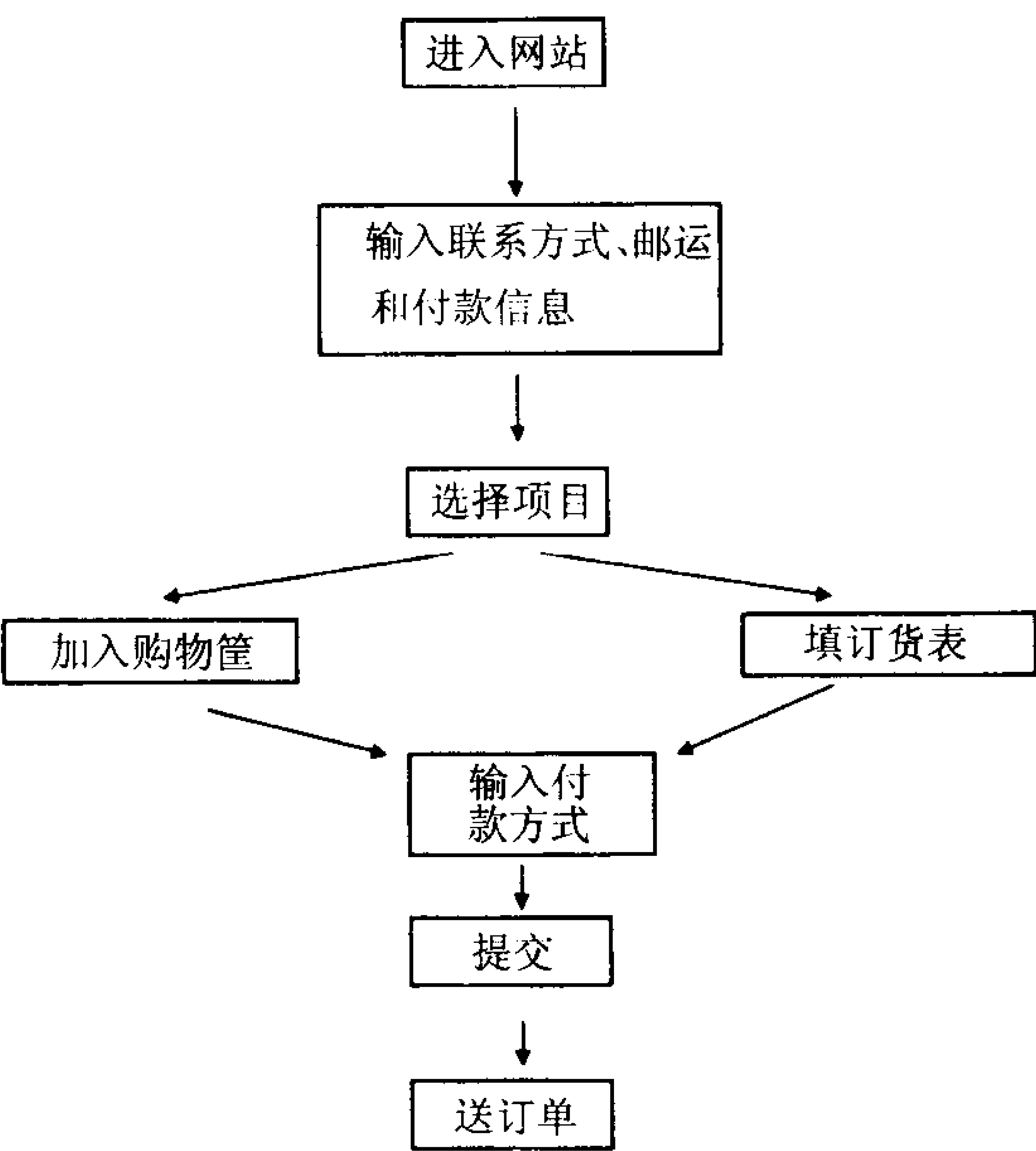


图 21.3 网上订货

(1) 选定拥有你想要产品的网站，该网站可以是一个供应商的网址，市场的或你们组织的网络目录。

(2) 从某方面说，不管你在哪，只要你想要网上购买产品，你就需要注册，并提交联系方式，邮购和付款的相关信息。这些信息可以以“家常小甜饼”的方式存在你的计算机里。如果你改变或删除这个文件，或要在其他的地方进入该网站，你就需要再登记，输入一次。如果你已登记，进入想进的网站就要再输入你所委托的人或选择的用户姓名和密码，便可进入并订购你所想要的。以前的信息将自动补充上，只是大概的信息，以减少错误和查询中的混乱。

(3) 下一步，寻找或浏览所要的产品的网址，选择其上各项条款并订购。各网站一般使用运货车或使用在线订货单的方式。要注意的是并非所有的网站都使

用运货车，而且当使用者离线并一会才返回时，就不能再次使用了。

(4) 一旦各条款已填写在运货车上或是订货单上，你便可再调整数量，增加或删除某些条款。

(5) 如果该网站用运货车，那么，下一步骤就提交这个运货车，这个步骤就是订购。因网站而异，还可能再有一次检查的步骤，这也是你改变运货车上信息的最后一次机会。然而也并非都是这种情况，所以不要点击过快。如果你是在ATCG 目录中，各项条款可添加到多个供应商。这种类型的系统一旦运货车交上去，所选择的购物条款就交到了供应商手里，并在供应商的供货单中占有一席之地。如果该网站用订货单，那么就交订货单。

(6) 无论是在运货车还是订货单被提交前后，订货系统都会询问你的付款方式，一般是购买卡或采办卡的卡号。

(7) 当你填完并提交了信息，订货已成功。运货车或订货单就会又一次变空。

如果使用的系统是当地或集中的服务系统，它就会把你的信息用于组织内的资金申请。然后，当购物单提交上后订购就会按提前预订好的路径进行。管理者规定了被批准的途径，那些不被批准的可能是超出了钱数的限制。此时，将有一个你所订购的产品正待批准的 e-mail 发送到你的邮箱中，只是等待批准。这个信息会在网上通知，所以使用者应注意网上关于订购的信息情形。一旦通过批准，订购单就会发送给供应商完成订购，供应商也会将订购完成的这个信息直接发送给使用者或某个组织。在组织系统内，一旦订购成功，网络可以启动顾客系统使每个人的信息显示在网上。

21.5 结论

万维网正改变着今天的商业活动，并且在将来还会改变。繁琐的没有价值的程序正在被削减，正如各组织正努力对其内部进行调整，以致降低成本，保持竞争力。

如今，电子市场的追随者们认为，电子市场是获得信息方式和入口的趋势。这并不意味着供应商的网站的灭绝，而意味着其内容的不断变化和逐渐适应，所以各网站仍很重要。如果真是未来提出的信息成为使用者的选择的话，让我们期待有更多固定的网站能从事有关化学、分子生物学材料等的这一类特定产品的工作吧！

(吕文发　译)

第四部分　计算机和分子生物学：信息发布与限制

22 网络计算

——重塑科学家使用计算机的方法和
人们对计算机的观念

Brian Fristensky

22.1 引言

本文描述的网络计算机(network computer，NC)是一种相对独立的 PC 机。通过转移数据和对服务器的处理，使每个用户都可通过 NC 完成所有工作。NC 向所有用户提供可信持续的接口，并使之很容易就获得如实验数据库等资源。NC 具有不会过时的特点，通过共享硬件、软件和管理方法而节约开支。将来，PC 机驱动的实验室设备可能被以 Java 为基础并受到网络监控的 NC 机器人所取代。

22.1.1 问题：以往的奢侈型客户

独立的 PC 机被认为是“奢侈型客户”，因为它必须具备完成全部任务所必需的软件和硬件。随着内存的飞速发展，促使软件要相应升级才能利用这些内存。因而大多数 PC 机和与之相匹配的硬件和软件必须每 3~5 年更新一次。

系统管理是 PC 机中最大的隐性消耗，因为每台 PC 机都要根据不同用户的不同目的而进行配置。PC 机变得越来越复杂，尤其是因为它们与网络的结合。甚至对专业的 PC/LAN 管理者来说保持工作中的每台设备运行和顺利升级也成为一个越来越不现实的目标。

大家都知道当仅有的安装了所需程序的机器被占用时，就必须等待，这极为不方便。而且每台 PC 机都是专用设备。举个例子，为了在文件中进行序列排列，你首先要在实验室的 PC 机上运行，那里有所必需的序列分析程序，然后上传到局域网(LAN)目录下，再进入办公室，下载该文件，再把它放入文字处理器中。当你决定要以略微不同的方式改变这个序列时，你就不得不重复上述操作。

PC 机中的软件和数据碎片会浪费时间，并且使你难以记住特定文件的位置，或是最近使用的版本文件的位置。而且由于共享的 PC 机很少做备份，以 PC 机为基础的操作系统几乎完全缺乏安全性，其他用户能够通过非法或恶意的手段毁坏有价值的数据。

22.1.2 解决方法：未来的节俭型客户

早期计算机是集中管理，通过远程的终端以纯文本的方式服务于大量用户。今天，像 UNIX 这样的系统能向几十个或上百个用户提供点对点的桌面(point-and-click desktop)服务。有几种类型的网络计算机正在发展之中。在 20 世纪 80 年代中期，X11 或 X-Windows 系统已在 MIT 方面领先，现在得到 Open Group[1] 的发展，它是最稳定和最广泛使用的协议。如图 22.1 所示，用户连接到注册服务器，该服务器把 X11 命令发送给 NC 机，X11 命令使每个用户窗口的内容和位置个性化。最典型的例子，执行大多数程序需要移动窗口，像滚动、剪切、粘贴、打开或关闭窗口，X11 卸载这些任务到 NC，减少服务器的负担。而所有其他任务由服务器完成，所以 NC 被认为是节俭型用户。X-终端不能运行光驱，不能自动运行应用程序。

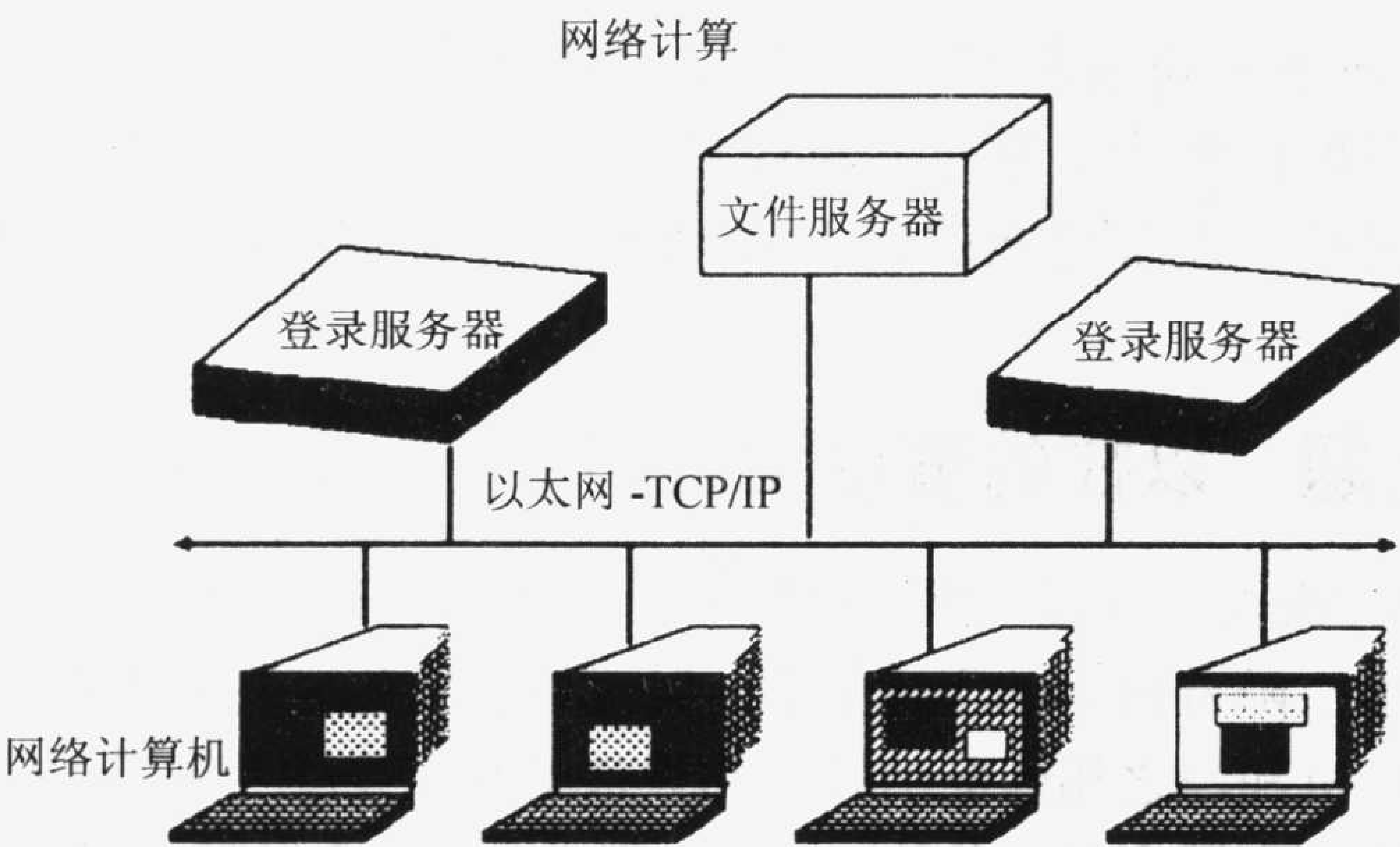


图 22.1　网络计算：数据和软件占据中心文件服务器，固定在所有远程服务器上。任何用户都能从任何 NC 注册进入服务器并完成任何任务

当可获得多项服务时，所有软件和文件集中在文件服务器，该服务器的文件系统通过远程网络添加到所有注册服务器上。无论从哪个服务器登陆，都会获得相同的目录和界面。简单地说，网络计算的核心概念是任何用户可在任何 NC 中完成任何任务。

图 22.2 表明的是在 Manitoba 大学用我们的 Sun/UNIX 系统用户所能获得的功能。在顶部，是用 WordPerfect 为 UNIX 写的授权决议[2]。寡核苷酸背景信息是从我们的 AceDB[3] 上获得的驱动的实验室数据库得到的。在项目中使用的关于基因的信息是使用网络代理 Nentrez 访问 GenBank 获得的[4]，后者在背景的右下部。在左侧较低处，程序名为“MS-Windows 程序管理员”是运行 Sun 公司的 Wabi[5]，用 JetForm 授权的软件填写表格。最后，需要上网查找信息时，可通过点击图标

打开网页使用浏览器进行网页浏览。

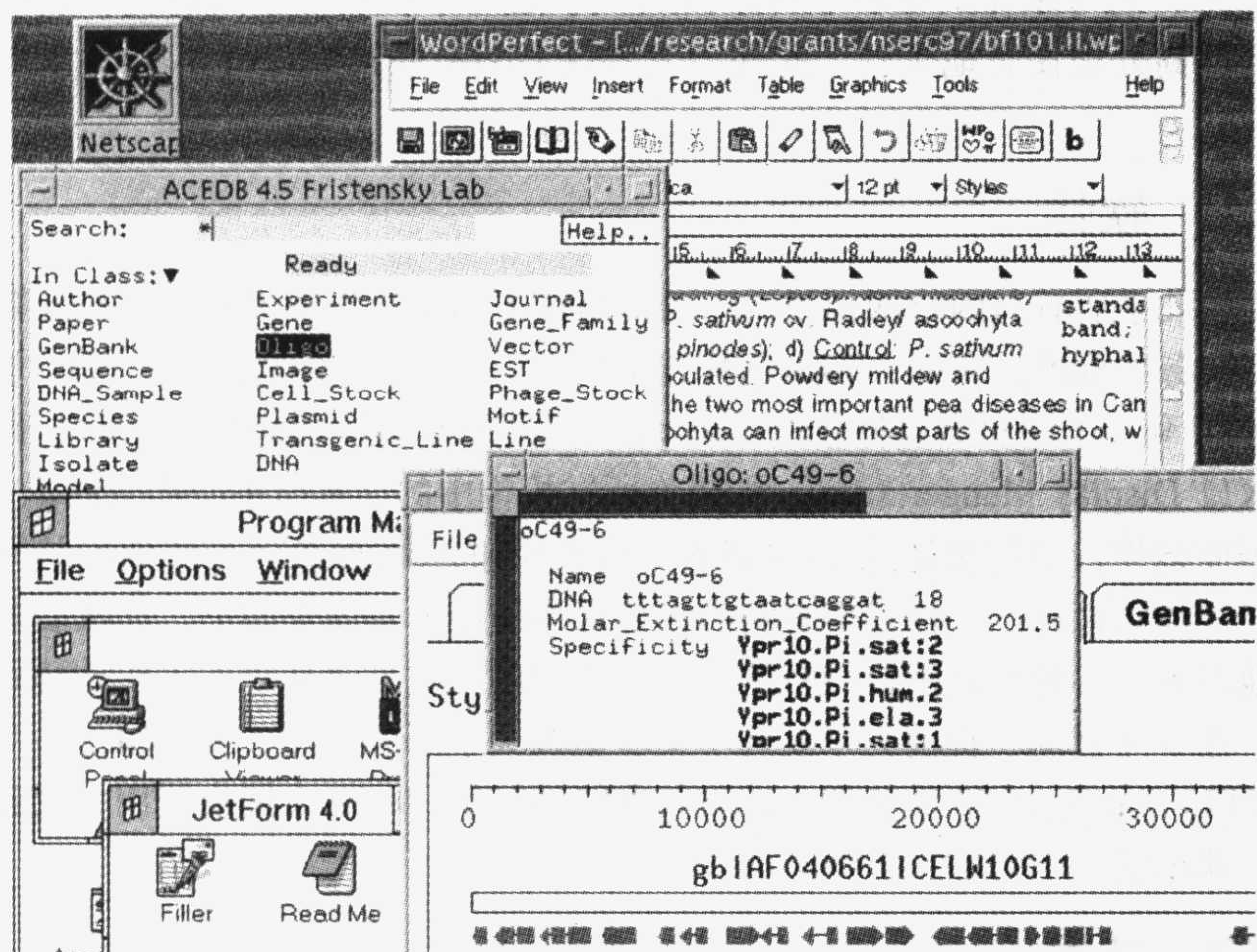


图 22.2　典型 UNIX 会议的屏景，显示了大范围的桌面和可得的科技软件

在这种系统中，软件通常由系统管理员来安装，但用户也可以安装或书写自己的软件。我们大学的序列分析设备是基于没有根目录特权的一个普通用户账号得以发展和管理的[6]。

我想指出的要点是：在 NC 机上可以完成所有通常在 PC 机上完成的任务。自 1990 年以来，我专用 UNIX 来进行计算，而我的实验室中甚至没有 PC 机，我们在 1993 年购买的第一个 X-终端仍能使我们在服务器上运行最新的软件，还可以很容易升级操作系统和应用软件，甚至升级到 64 位服务器而不需要重新构建终端。

22.2　硬件、软件和技术

22.2.1　硬件

需要一个以 UNIX 为基础的工作站，如 Sun/Solaris、IBM/AIX 或 Intel/Linux，不需购买 32 位工作站，UNIX 的大多数版本都完全适用于 64 位体系结构，或者在几年以后会适用。如果你所在的机构提供了中央化的 UNIX 服务器，你甚至不需购买服务器。同时你还需要一个 X-终端，或运行 X-Windows(模仿)软件(见

22.3.5 节)的 PC 机。确保任务顺利完成的一个重要的因素是 RAM。服务器的存储器越大，它与磁盘交换程序和数据的频率越低，额外的存储器通常能从因特尔的经销商处很便宜地买到。此外，大多数 UNIX 版本能支持 2 个或 2 个以上的 CPU。

22.2.2 软件

由于几乎所有与因特网相关的东西都是在 UNIX 下发明的，典型的配置达到了全网服务的标准，包括 telnet(远程登录)程序、ftp(文件传输协议)、e-mail(电子邮件)程序和一个网页浏览器(如 Netscape)。典型的 UNIX 系统也包括服务于 X-终端的 X11 Display Manger(xdm)。如果你所在机构中的计算机中心作为网络 UNIX 服务器而运作，他们可能乐于(可能需少许费用)把你的服务器当作他们的终端计算机之一加以管理。这样是很有利的，如果你的服务器需要卸下，你能登录到任何其他服务器。举一个例子，在 1996 年我把我的主页目录从个人工作站移到校园系统，在 3 个月间，提前为学院几个实验室购买了新服务器，我使用了大众认可的双 CPU Sun Sparc20 服务器，该服务器能同时供 30~60 个用户使用，我极少见到服务器出错。

22.2.3 技术

如果你的计算机中心按如上所述管理你的服务器，你将获得 NC 模式的最大利益。让专家来做这项工作，这对安全尤为重要，大部分专业的管理系统拥有填补最新安全漏洞补丁和实时适当安装的人。换句话说，在你的实验室或学院中，需要有人具备使用 UNIX、一些编程经验(最好是 C 语言或 Java)和制作 HTML 的能力。其他人简单入门就可以受益于 UNIX[7]，但从总体上讲，学会基于 Windows 系统的相关操作者将易于向 UNIX 界面转移。

22.3 应用事件

网络计算机仍在向前发展，并且可能会出现许多不同的节俭型客户模式的应用程序。由于在近 10 年里 UNIX/X-Windows 对网络计算机的改善已经相当稳定，并且可能成为一个主要的网络计算机模式，我将讨论一些你想知道的关于 X-Windows 平台所做的计算。

22.3.1 选择最佳方案

当你移入 NC 平台，就应当始终如一。如果在 NC 和 PC 间分开你的计算和数据文件，会使事情变得复杂。但如果始终使用 NC 平台，你会发现基于服务器的计算方式会更快。

22.3.2 第三方软件问题

网络计算发展的主要障碍是缺乏以服务器为基础平台的第三方软件，从一定程度上讲，大多数的桌面软件是专门为 PC 机创作的。由于 Java 的使用可能导致在未来平台的独立，所以目前更难以找到针对服务器为基础的系统，如 UNIX、VMS 或 AS400 所制作的应用程序。另一方面令人惊奇的是这些平台可利用的软件的数量如此之多，在许多情况下，可以使用以服务器为基础的编程版本，如 WordPerfect[2]或 Adobe PhotoShop[8]。在其他情况下，也可以使用专门为用户/服务器平台设计的可比较的应用程序。

22.3.3 在 UNIX 下使用 Windows 应用程序

一般情况下，最好尽可能地使用本地 UNIX 工具。如果得不到 UNIX 版本，Windows 应用程序能通过 NCD's Wincentre[9]从 WindowsNT 服务器上显示到 X-Windows 桌面上。这种解决办法的有利因素是：这些应用程序可在本地 Intel 体系结构上运行，并把一些任务下载到 NT 服务器上；其不利因素是：需要有指定的 NT 服务器，并且 NT 服务器必须配置成与 UNIX 文件服务器同步运行。对于 Sun 系统，Wabi[5]能在软件模拟器中运行 Windows 3.1。

22.3.4 来自远程服务器的 X11 程序

如果能在另一个服务器上而不是你近期登陆的服务器上运行 X-Windows 应用程序，将会是件令人惬意的事。例如，当一个应用程序可能在该服务器上不能运行时，而登录进入有运行许可的服务器和对终端或工作站设置了 X11 显示时，这个问题就很容易解决了。

如果名为 raven 的用户登录进入 marigold. uofm. ca，但想运行安装在 Petunia. uofm. ca 的 SAS，进入 petunia 使用 telnet：

```
{marigold:/home/raven}telnet petunia
Trying 130.122.36.48...
Connected to petunia.uofm.ca.
UNIX(r) System V Release 4.0(petunia)(pts/18)
login:raven
Password:
{petunia:/home/raven}
```

然后，对终端或工作站设置环境变量 DISPLAY。

```
{petunia:/home/raven}setenv DISPLAY ncd12.uofm.ca:0.0
```

这个命令会使所有接下来的 X11 程序进入命令解释器的位置而显示在称为 ncd12 的 X-终端屏幕上。

最后，进入 SAS：

```
{petunia:/home/raven}sas &
```

注意：这个命令后的&符号必须写上。在每个 UNIX 命令后放一个&的符号，从而使之在后台运行。这样就做了两件事，第一，它释放了命令窗口而在这个应用程序运行时你可以做别的事；第二，如果你愿意，你可在不终止 SAS 时退出 petunia。

对于那些必须经常运行在远程服务器上的应用程序，这些步骤可通过把这些命令放入一个命令行解释器而使之自动运行，这些指令可从工作站菜单中启动。

22.3.5 使 PC 变成 NC

许多实验室都在 PC 机硬件方面做了大量的投资，同样大多数部门还有不能运行最新软件的旧 PC 机(如不能运行 Windows'95 的 486)，最后，一次性购买大量的 NC 的做法通常是不可取的。由于上述原因，一些 MS Windows 和/或以 Mac 为基础的程序允许一台 PC 机像 X-终端一样运行。

一些 X11 模仿软件的优点包括：

- 价格低廉。
- 软件经长时间应用证明是可以信赖的，并易于使用和安装。
- 即使是 486 也能像 X-终端一样快速运行。并且移动窗口的命令基本相同，所以一旦有一台旧的 468 工作了，它会运作得很好。
- 包括了网络传输协议，如 SLTP 和 PPP，能够在家里通过快速 modem 运行 X-Windows 任务。

为什么 X 模仿软件是中间产物而不是永久的解决方法？有以下几点原因：

- 要求保持 MS Windows 的运行。否则，每次在 PC 机或 LAN 上升级 Windows 或安装软件，或更换 PC 机的网络软件，都可能影响 X 模拟程序。
- X 模拟器是不完美的。由于 PC 机和 Windows 增加了复杂的层，X-终端并不支持 X11 软件执行的所有程序。
- 如果你正在使用带有 X 模拟器的 PC 机，做一些事时就要在 X 桌面和 PC 机上来回切换。如此一来一些文件就要被破坏，又必须上载和下载一些文件，这样你就很少有兴趣去真正学习使用 X 桌面。但如果只使用一个系统，事件就简单得多了。

解决方法是在短期内使现存的 PC 机升级到 X 模拟，或为长远打算而购买新的 X-终端而不是新的 PC 机。PC X 模拟程序包括 Hummingbird's eXceed[10]、White

Pine Software's eXodous[11]和 NCD's PC-Xware[12]。许多此类经销商提供免费下载程序的试用版。

22.4 未来：NC 将如何改变我们的工作方式

此部分将展示 PC 机在网络普及时的发展前景，我支持使用已发展完善或正在向前发展的事物。但 22.4.4 节是个例外，它是这些发展的一个综合。

22.4.1 无论何处都以相同的方式运行

当今，每个研究者和学生都使用独立的 PC 机，甚至使用几台专门为不同目的而设立的 PC 机。我们不得不把文件拷到磁盘上才能在家或我们随身携带的笔记本电脑上工作。随着 NC 出现在旅店、机场和大学的计算机中心和图书馆，在不久的将来你就能通过 NC 在家里进行工作，进行高强度的计算，甚至检查实验进程(见 22.4.4 节)。由于 NC 便宜，它们将遍地皆是。在机场停留期间，在旅店房间里或假日旅行时，你都可以像在实验室或办公室一样工作，你将不再受到携带文件软件和电源的限制。

为了在旅行时简化 NC 的使用，一些公司，如 Network Computer Inc[13]已经制作了一些智能卡，这些智能卡携带了可通过因特网找到主页服务器并连接到用户所需账户的信息。

22.4.2 电子研讨会、宣讲、授课

课堂和座谈会的形式被代之以网页，然而与 NC 比较，网页浏览器仍然受到限制。例如，我使用 X-终端，X-终端屏幕上的内容发送到一台 1024×768 的投影仪上，投影出我的细胞遗传学课程的讲义，然而大部分讲义是在网页[14]上，我常使用图形工具来做个简单的演示。以服务器为基础的终端能确保讲解能像在办公室一样，研讨会发言者将从他们的主页服务器上得到在课堂上所能得到的图像、数字和程序，甚至与问题相关的其他数据或数字也会显示出来。

22.4.3 Java：一次性写入，随处运行

Java[15]是来自 Sun Microsystems 的新的编程语音，是专门为在所有计算机平台上运行而设计的。其工作方式为：Java 程序在被称为 Java 模拟机的命令解释器中运行，由于 Java 模拟机已被装入几乎所有的计算机平台(如 UNIX、Windows、OpenVMS、IBM mainframes、Macintosh 等)，所有的 Java 程序能在各种平台上运行。你所需要的是 Java 模拟机，因此用 Java 写的软件是独立平台，软件开发者只需书写并且维护一个 Java 软件版本，而不是为多种平台制作多种版本。

Java Molecular Biology Workbench[16]是一套 Java 程序的例子。在这种情况下，

程序以 applets 的形式运行，该程序是从远程的服务器上下载的。Java 程序也能在本地服务器、工作站或 PC 机上作为独立的应用程序被下载和运行。

Java 的一个主要优势是它是模块化的，用传统语言写的程序是每一个整体占用数百万字节的内存。Java 应用程序是为专一目的设计的，每个软件包执行单一的功能，一台 NC 运行一个 Java 应用程序时，只需将所需的目标在指定时间内从服务器上下载即可。以 Java 为基础的 NC 不需大量的内存和处理能力，后者提供的这种保护已过时了。

22.4.4 消除计算平台、网络、实验室笔记本、实验设备的合并

最初 Java 被认为是依赖于硬件设备的语言，这使得网络实验设备的多功能性和可升级性要好于我们当前的实验室设备。今天，实验设备，如荧光摄像机、DNA 测序仪，甚至植物生长的温度都需一台 PC 机来操作。由于分析和数据获取的软件都会占用指定的 PC 机，你就不能在设备被占用时分析你的数据。在将来许多设备将是网络的 NC，一个明显的结果是：不再需要给每台设备购买和安装 PC 机，指定的硬件，如监视器、打印机或 LCD 显示器将也不需要，这使得 Java 设备更小更便宜，拥有一个 Java 芯片[17]的设备将完成所有操作。对每台设备可通过任何 NC 进行操作和监控，在实验完成后，数据直接加载到用户的目录以便分析(图 22.3)。

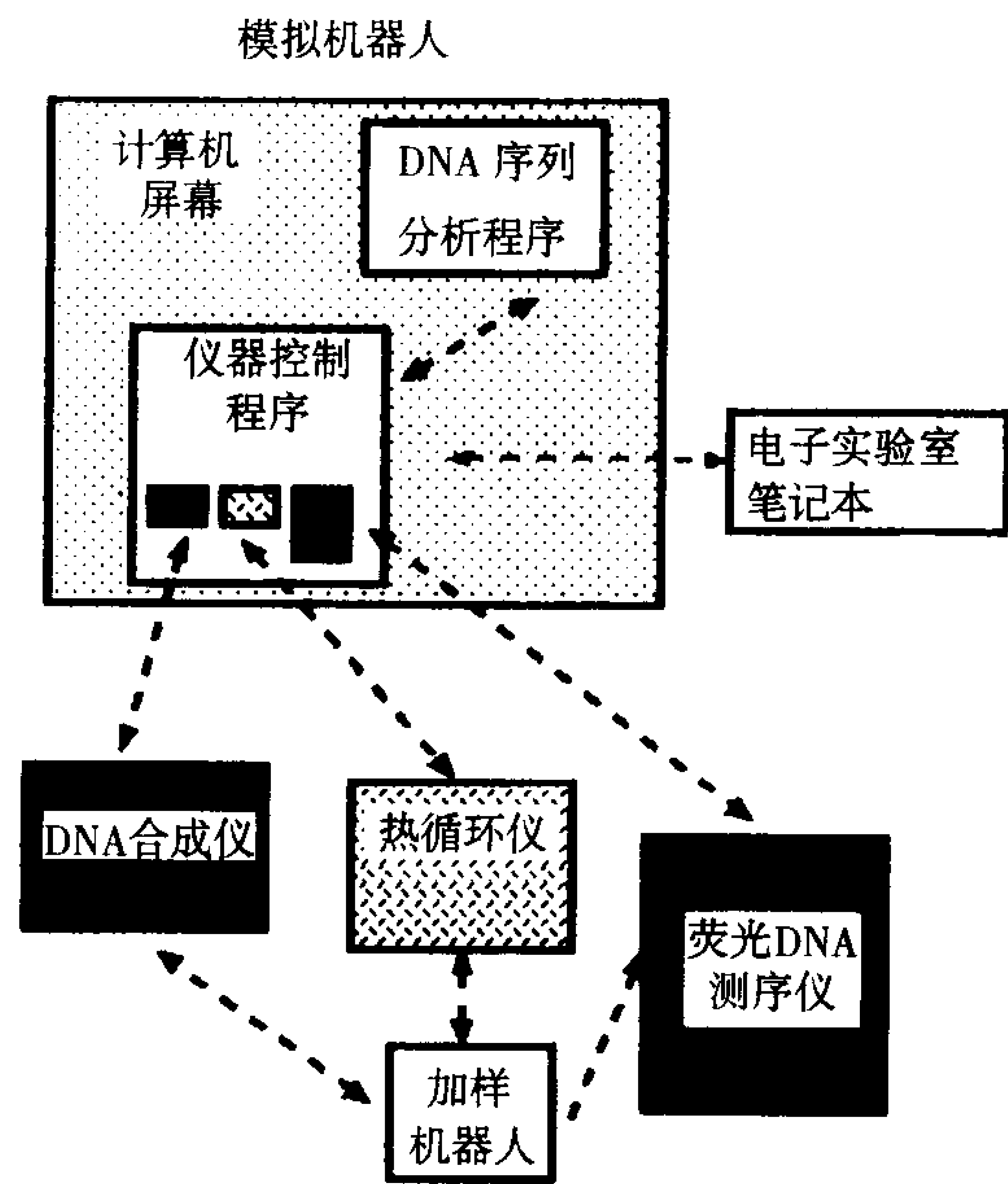


图 22.3 模拟机器人

Java 编程的网络设备能通过软件连接在一起，创建虚拟机器人使其适应具体任务，并受 NC 控制

更令人兴奋的是虚拟机器人的概念，从理论上来说，虚拟机器人可通过与仪器控制程序和 Java 设备连接而创建在桌面上。在该例中，仪器控制程序通过屏幕图标再现每个真实的设备。注意到当一台移液机器人被每台设备呼叫时，不需要在控制程序中重复，这就近似于面向对象编程语言(如 Java)中类的继承。在图 22.3 中，一个 DNA 序列分析程序用来设计引物，它被送到设备控制程序中，一台 DNA 测序仪机器人是通过依次连接 DNA 合成仪、热循环机和 DNA 测序仪而形成的。测序结果通过控制程序从测序仪传送到 DNA 序列分析程序。电子实验室记录本也是 Java 设备，它能用来确定用于 DNA 测序的样品或热循环机产生的样品存放位置的程序。

由于使用现在 PC 和操作系统可以创建多功能机器人，每台虚拟设备是需扩展编程来完成的专用设备。一个以软件为基础的控制程序能使虚拟机器人适合于每个任务。例如，如果你想在装入测序仪之前知道你的 PCR 产物的量，荧光探测器可以连接热循环仪和测序仪，就会被检测到成功扩增的样品。

22.5 远景及展望

当我们在 5 年或 10 年后回顾 PC 时代，我会为过去我们在今天认为理所当然的事而惊奇。最荒谬的是这样一个观念：每个用户都是管理者。在 NC 时代，服务器接受专门的管理。就用户而言没有必要来管理 NC，他们所做的只是使用 NC。

计算机将更加经济实用，从而打破过去的恶性循环，用户将继续把金钱用于计算，而不是用于购买 RAM、光驱、协处理器和软件，我们将向那些给我们提供这些事物的服务商支付月租，专业的维护、集中的资源将更稳定更可信。

今天，大部分用户习惯于 MS Windows 操作系统。在 NC 时代，界面和操作系统将会竞争激烈，最终结果是用户有更多的选择和竞价。本文集中论述 UNIX，主要是因为 UNIX 特别适于网络计算机。然而，NC 模式的开放意味着其他系统，如 OpenVMS 和 AS/400，也可能是 Windows NT(如果可测量性和安全问题能被解决的话)在提供 NC 服务中发挥一定作用，可能的结果是 NC 服务提供者将使用不同服务器和操作系统的混合搭配在单一界面来传递全程应用程序，直接传输到用户。无论服务器终端如何改变，用户对 NC 硬件的投资都将回馈，因为客户越偏向节俭型，它就越不易过时。

对于 PC 时代的回顾，在 DEC 的主席和创立者 Ken Olsen 1977 年发表的讲话中可见一斑：“谁也没有理由想要计算机呆在他们家里(或实验室，B. F.)”。

关于网络计算的重多信息请查阅：`http://home.cc.umanitoba.ca/~psgendb/nc`。

(欧阳松应　译)

参 考 文 献

[1] The Open Group (1998) *X Window System,* http://www.camb.opengroup.org/.

[2] Corel *WordPerfect* forUNIX, http://www.corel.com/products/cwp7unix.htm.

[3] Durbin, R. and Thierry-Meig, J. (1991) A C. elegans Database. Code and data available from anonymous FTP servers at lirmm.lirmm.fr, cele. mrc-lmb.cam.ac.uk and ncbi.nlm.nih.gov.

[4] Natl. Center for Biotech. Info. Nentrez, http://www3.ncbi.n1m.nih.gov/Entrez/Network/nentrez/overview. html.

[5] Sun Microsystems, *Wabi,* http://www.sun.com/solaris/wabi/.

[6] Fristensky, B. (1999) Building a multiuser sequence analysis facility using freeware, this volume, pp. 131-198.

[7] Sobell, Mark G. (1995) A Practical Guide to the UNIX System. Addison-Wesley Publishers, Reading, MA.

[8] Adobe Corp. *PhotoShop.*http://www.adobe.com.

[9] Network Computing Devices Wincenter. http://www.ncd.com/pwin/pwin.html.

[10] Hummingbird Ltd., http:/www.hummingbird.com/products/exceed/.

[11] White Pine Software, http://www.wpine.com/exodus/.

[12] Network Computing Devices, http://www.ncd.com/ppcx/ppcx.html.

[13] Network Computer Inc. http://www.nc.com/prodcard.html.

[14] Fristensky, B. *Introductory Cytogenetics,* Univ. of Manitoba. http://www.umanitoba.ca/afs/plant_ science/ COURSES/CYTO/.

[15] Sun Microsystems. *The Java Platform.* http://java.sun.com/aboutJava/.

[16] Toldo, L. (1997) *JaMBW 1.1: Java-based molecular biologists' workbench. Comput. Appl. Biosci.* **13**, 475-476. http://www.embl-heidelberg.de/~toldo/JaMBW.html.

[17] Sun Microsystems. *Java Computing.* http://www.sun.com/java/.

23 利用DNA进行计算

Lila Kari   Laura F. Landweber

23.1 计算学历史上的新参与者

简要回顾一下人类的历史我们就会发现，原始社会时人类就需要进行数数和计算，或度量月份和季节，或用于商业和建筑。用于计算的手段多种多样，任何可能的形式都可用于计算，因此逐渐出现了手工计算、机械计算(算盘、机械加法器)，而且，后来出现了电子计算设备。在人类进行计算的历史长河中，一直努力应用最好的计算技术，电子计算机就是最新的计算技术。尽管在50年前计算机的出现给计算科学带来了一次革命，但它并不代表着计算科学发展到了顶峰。实际上，电子计算机也有其局限性：它所能存储数据的容量有限，而且按照物理原理，电子计算机的运算速度马上要达到极限。最近，为突破这个障碍，人们尝试再一次更换计算工具：用生物学工具替换电子元件。

DNA计算(有时也称为生物分子计算或分子计算)是一种新的计算范例，它应用生物分子解决计算问题，同时应用自然的生物过程作为计算模型。Leonard Adleman所做的一个实验开创了这个领域的研究，他在1994年应用分子生物学的工具解决了一个复杂的计算问题[1]，这令学术界感到很惊奇。他的实验通过操作DNA链解决了一个定向哈密尔顿路径问题(Hamiltonian path problem)的实例。这是人类首次应用生物学来解决数学问题。

利用生物分子(主要是DNA)进行计算提供了一种用于执行和观察计算过程的新方法，这是一件令人异常兴奋的事情。其主要的思想是用DNA链代表数据，应用分子生物学工具执行计算操作[1a]，这种方法不仅新颖，而且分子计算有超越电子计算机的潜能。例如，DNA计算能用比电子计算少十亿倍的能量，然而能节省千亿倍的空间[2]。而且利用DNA进行计算具有高通量的优点，理论上成万上亿的DNA分子可以同时进行化学反应，也就是同时进行计算[3]。

尽管这项技术很复杂，但是DNA计算所蕴含的思想源于一个关于两个过程的简单类比，这两个过程一个是生物学的，一个是计算科学的：

- 从根本上来说，一个有生命的有机体的复杂结构源于对DNA所编码的信息所进行的一系列简单的操作(如复制、剪切、插入、缺失等)。
- 任何计算，无论多么复杂，都是非常简单的基本算术和逻辑运算组合的结果。

Adleman 意识到这两个过程不仅相似，而且应用现代分子生物学技术可以做到用生物学来解决数学问题。更精确一点说，DNA 链蕴含着信息，而且各种分子生物学实验技术可执行简单的操作。(读者可阅读参考文献[4]，以获得更多的分子生物学概念。)DNA 序列编码信息而且能够对 DNA 链进行简单的操作，使 Adleman 解决了一个 7 节点的引向哈密尔顿路径问题[1]。

当一个导游图 G(带有设计好的结点 v_{in} 和 v_{out})存在一个单行线顺序(e_1，e_2，e_3, …, e_z)，可以实现从 v_{in} 开始到 v_{out} 结束，并且每个结点只访问一次时，这个导游图 G 就可称为存在一个哈密尔顿路径。大家所知道的“旅行商人”的问题是这个问题的一个简化版本，它提出了下列问题：任意给出一组城市，需要商人游历(图 23.1)，遍历这些城市最短的路线是哪条？Adleman 版本通过给定初始和结束的城市限制了路线的数目。因为所有的城市都没有联系，关键的任务就是发现一条(如果存在的话)连续的路线能把它们都连起来。

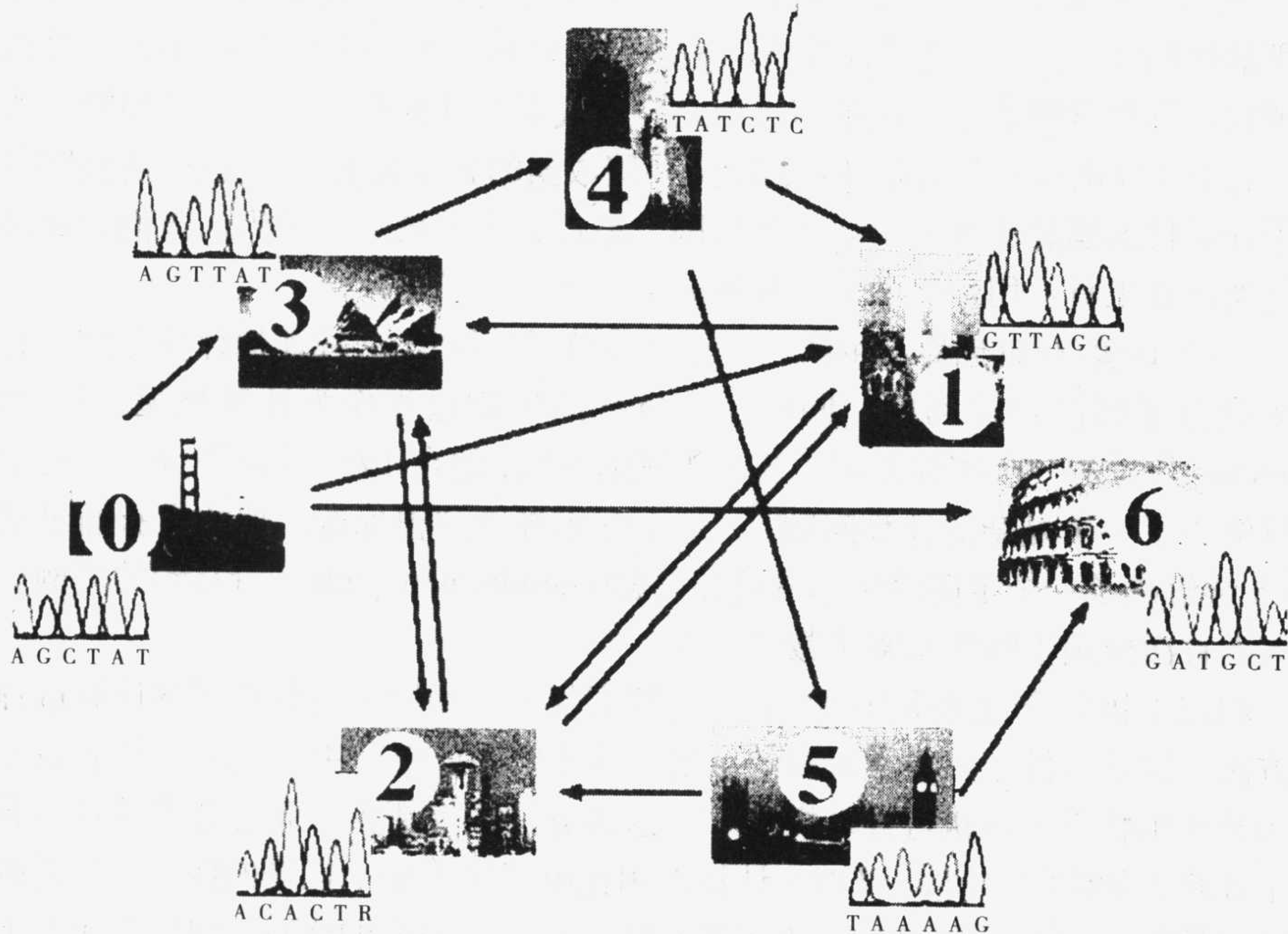


图 23.1 Adleman 实验[1]中所用到的一个实例图

城市即结点，用任意的 DNA 序列来代表。旅行商必须找到能走遍 7 个城市的最短的路径。

在这个例子中，开始于旧金山(城市 0)，最后到达罗马(城市 6)

下面是解决这个问题的算法：

(1) 由图随机生成路线。

(2) 仅留下从 v_{in} 开始到 v_{out} 结束的路线。

(3) 如果图中含有 n 个结点，那么仅保留那些恰好通过 n 个结点的路线。

(4) 保留那些一次性通过图中所有结点的路线。

(5) 如果还有路线存在，则存在符合条件的路线，反之则无。

执行步骤(1)时，图中每个结点被编码为一条随机的含有 20bp 的 DNA 链(或称寡核苷酸)。接着，对于图中的每一条边(带方向)，生成一个不同的 20bp 的寡核苷酸，它包括源结点序列后一半的互补序列，再加上目标结点的前一半序列。以这些互补的 DNA 寡核苷酸作为“夹板”，所有的对应于合适边的 DNA 序列将自动组装并通过 T4 DNA 连接酶连接起来。因此，退火和连接反应生成了包含遍历图中结点的随机路线的 DNA 分子(图 23.2)。

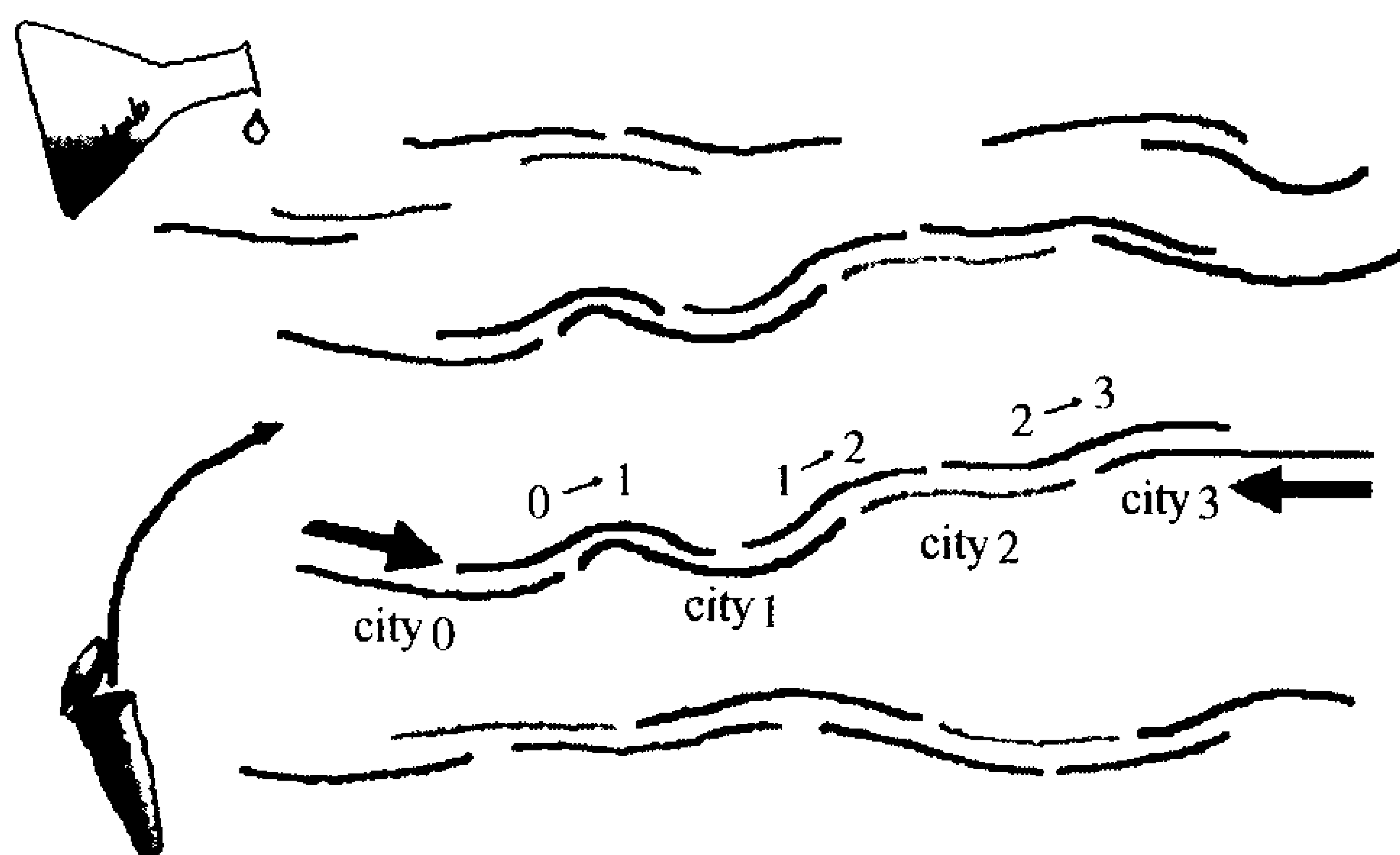


图 23.2 代表路径的 DNA 分子的自组装

用 PCR 引物标记开始及末端的寡核苷酸序列(图中是城市 0 和 3)，图中用箭头表示。

第 i 个城市的下游序列和 $i \to j$ 这条边的上游序列互补，同样 $i \to j$ 这条边的下游序列第 j 个城市的上游序列互补

执行步骤(2)时，步骤(1)所得到的产物通过聚合酶链反应(PCR)进行扩增，引物使用代表 v_{in} 和 v_{out} 的序列。这个扩增过程得到的产物是仅包含从 v_{in} 开始到 v_{out} 结束的路径。

执行步骤(3)时，利用琼脂糖凝胶电泳分离并回收长度正确的 DNA 条带。如果存在所需要的路径，它将通过所有 7 个点，每个点对应的序列长度为 20bp，因此其长度应该是 7×20 = 140bp。

步骤(4)通过连续应用亲和纯化法对除始末结点外的每个结点对应序列进行纯化来完成。这个过程实现了在不纯的条带中分离并回收包含给定结点的单链。与结点序列互补的 DNA 条带将会吸附到磁珠上。包含单链 DNA 的不纯的溶液通过磁珠时，包含结点序列的条带将会被选择性地保留。缺乏任一所需结点序列的条带在步骤(4)之后将不复存在，因为它们至少在一次过柱时被冲走，而没有被保留。

执行步骤(5)时，用 PCR 会检测到包含哈密尔顿路径的分子。第一次 PCR 扩

增步骤(4)的结果，并用类似步骤(2)的方法检测产物是否存在。如果存在，再进行第二次 PCR，用与每个结点序列互补的 DNA 寡核苷酸作为引物，目的是检测上次 PCR 产物的内部结点。这个步骤不需要进行 DNA 测序便能绘制出各个结点连接成的序列图谱。

Adleman 实验结果的一个著名的发现不仅解决了这个数学问题，而且从下面所叙述的意义上说，它也是一个复杂的计算问题[5, 6]。

问题的难度等级可以根据用最优的算法在一台计算机上解决这个问题所用的时间来衡量。算法的时间复杂函数用一个多项式函数来限定，根据所描述问题输入值的大小，算法的时间复杂度用多项式 P 来代表，这样的算法一般情况下效率较高。任何算法的时间复杂度函数都不会限定于一个低效的期望值 EXP。如果一个问题特别难，没有一种多项式算法可以解决，那么这个问题就叫作“难以处理”的问题。

一个特殊类型的问题，看起来是难以处理的，它包含 P 但包含于 EXP，它叫作“不确定的多项式时间”，或叫作 NP。下面是问题类型的包含关系式链：

$$\mathrm{P} \subseteq \mathrm{NP} \subseteq \mathrm{EXP} \subseteq \text{Universal}$$

NP 包含这样的问题：没有已知的算法可以解决，但可能存在一个不确定的计算方法(这个方法有能力能一次性执行无数个独立的计算)。定向哈密尔顿问题是 NP 问题中的一个特殊类型，叫作“完全 NP”。一个完全 NP 问题具有这样的特性：其他的 NP 问题也能简化成它这样的多项式问题。因此，从某种意义上说，完全 NP 问题是 NP 问题中“最强硬”的。

关于“完全 NP 问题是否是难以处理的？”这个问题(数学上就是是否 P 等于 NP)，现在被认为是数学和计算科学中最前沿的问题。因为定向哈密尔顿问题看起来属于完全 NP 问题，好像没有高效的算法可以在计算机上将其解决。

Adleman 之后的其他实验也利用 DNA 操作解决了一些数学问题。Kaplan 等[7]重复了 Adleman 的实验；Guarnieri，Fliss 和 Bancroft 应用水平的链式反应扩增 DNA[8]；Wisconsin 计算机学家和生物学家学组研究出应用表面化学解决 5 变量的 SAT 问题的方法[9]；Quyang 等[10]应用限制性内切核酸酶解决了一个 6 变量的完全 NP 问题(Maximal Clique 问题)；我们其中一个实验室[11]最近用 RNA 解决了一个 9 变量的 SAT 问题。

与此同时，研究者们已对 DNA 计算在许多方面的可行性进行了大量研究：提出了 Adleman 问题效率较高的方法[12]；研究了 PCR 引起的复杂性[13]；研究了 DNA 自组装的应用[14]；DNA 结构[15]；报告了分子中连接和旋转的数据[16]；研究了利用 PCR 进行 DNA 的串联[16, 17]；在评估简单的 Boolean 公式方面也取得了一定进展[18]；引导 DNA 连接实验用于计算[19]；执行了一个基于分子计算的专门的工具“推论引擎”[20]；得到了解决最短的公共超级字符串问题的部分方法[21]。

理论方面的研究提出了许多利用 DNA 操作解决问题的潜在策略，这满足了 DNA 算法实验研究的需求。这些实验策略所描述的问题包括：SAT 问题[22]，破译数据密码标准[23, 24]，符号决定因素的扩充[25]，矩阵计算[26]，利用动态规划的算法解决图连接性及背包问题[27]，道路颜色问题[28]，超范围计算机代数问题[29]，限制性递归问题[30]和简单的 Horn 条款计算[31]。

23.2 向 DNA 计算机发展

目前提到的实验是单个的实验，可以建立模型解决特定的问题。这立即引出了两个基本的问题[1, 6]：哪些类型的问题可以用 DNA 计算的方法高效地解决？能否设计一个程序化的 DNA 计算机(至少在原理上)？虽然当前提出的 DNA 计算模型各不相同，但它们也有许多共同的特征。

实际上，任何类型的计算机，不管是机械的、电子的还是生物的，都需要具有两个基本的功能：存储信息和对存储的数据进行操作。下面我们提出两个问题：DNA 链中如何储存信息？哪些分子生物学技术可用于计算？为区别普通的数学运算和对 DNA 链的分子生物学操作，我们用“生物操作”这个术语代表后者。

一条 DNA 单链可以被看作一条包含 4 类不同字符 A、T、C、G 的字符串。在数学上，这意味着我们可以利用一个含 4 个字母的字母表($\Sigma = \{A, G, C, T\}$)来代表信息。顺便提一下，对于同一计算目的，这比电子计算机有更大的容量，因为电子计算机仅利用了两个数字：0 和 1。

关于对 DNA 进行的操作，所提出的 DNA 计算模型通常综合运用下列最基本的生物操作：

合成：通过合成得到一些不等长度的片段，适用于所有的模型。

混合：将两管成分混合，以得到联合成分[1, 32~36]。

变性：通过加热 DNA 溶液，将 DNA 双链解为互补的单链[35~39]。

退火(复性)：降低溶液温度，使 DNA 单链通过互补配对结合成双链[35~39]。

扩增(复制)：通过 PCR 反应复制 DNA 链[1, 25, 32~38, 40]。

分离：利用凝胶电泳或其他分离方法，根据长度大小不同，将 DNA 链分离[1, 32, 33, 36, 37, 40]。

提取：通过亲和性纯化得到含有某一子链的 DNA 链[1, 32, 34, 40]。

切割：利用 DNA 限制性内切核酸酶在特殊的位点切割 DNA 双链[37, 38, 40~42]。

连接：利用 DNA 连接酶连接两条含有相符黏性末端的 DNA 链[37~42]。

取代：利用位点特异性突变 PCR 反应，实现 DNA 链的替换、插入或缺失[40, 43]。

通过杂交标记单链：互补链相互结合形成双链(其实是对某一单链进行了标

记)，相反的操作是通过变性反应解除对单链的标记[9, 33, 35]。

通过多种核酶破坏标记的 DNA 链[9, 11]，或利用限制性内切核酸酶切割标记的链，并通过凝胶电泳对完整的链进行纯化[10, 33]。

探测和阅读：对于一管给定的成分，如果它包含至少一条 DNA 链符合所需操作的要求，并读出其序列，这样就得到肯定的答案，反之得到否定答案[1, 32~34, 36]。

一个生物计算包括针对试管中 DNA 链的一系列生物操作。这些生物操作可能是上面所列到的，也可能是其他的，这些生物操作被用于“编程”。一个程序得到一管 DNA(蕴含着信息)作为输入条件，返回“是”或“否”或者是一管新的溶液作为输出结果。

现在已经提出了许多基于上述生物操作的 DNA 计算模型，并且研究了它们的计算功能和可行性[1, 32, 37, 39, 43~51]。所提出的模型既有优点也有缺点，但总的来说，存在不同的模型，且各具特色，说明 DNA 计算的多功能性，而且更说明了制造一台 DNA 计算机器的实际可行性。

为制造一台 DNA 计算机，几乎每一步都面临着工程问题的挑战。最基本的一个问题就是在处理大规模系统和应对继发性错误方面所遇到的困难[52]。然而，我们注意到，有一些事件，生物体能自然地解决，如活性监测、调整生物分子的浓度、错误适应等：细胞必须调整不同成分的浓度以促进稀有分子的反应，它们也能处理自身存活不需要的产物。因为细胞能在体外成功地处理这些事件，这就暗示我们可以体外模仿这些过程。在这方面我们可以做理论上的假设，利用细胞膜隔离出一个空间，激活转运系统来有选择地通过其边界转运化学物质[53]。而且，家用电子计算机的设计原理可以应用到生物分子计算机的制造上[3, 54, 55]。

23.3 一个正式的 DNA 计算模型及其计算能力

DNA 计算方面的一项理论研究是寻找一个合适的正式的形式来描述分子计算。这方面的研究经常将分子计算模型的计算能力与一个图灵机(Turing machine)(当今电子计算机的正式模型)相比较。

我们通过一个插入、删除系统(DNA 计算模型的标准语言[43, 51])来描述这方面的研究。结果证明，这种 DNA 计算模型不仅在实验室是可行的，而且完全具有图灵机的计算能力。

在正式介绍这种模型之前，我们先总结一下它用到的专用术语[56]：对于一个集合Σ，“Σ ”表示集合中元素数目，即Σ 中元素个数。一个字母表是一个有限的非空集合。它的元素是字母或符号。字母经常用字母表中前面的字母来代表，有的有索引，有的没有，比如 a，b，C，D，a_i，b_j 等。(对于 DNA 计算来说，我们所需要处理的字母表是Σ $=\{A, C, G, T\}$)如果Σ $=\{a_1, a_2, \ldots, a_n\}$ 是一个

字母表，那么任意序列 $w = a_{i1}a_{i2}\ldots a_{ik}$，$(k \geqslant 0,\ a_{ij} \in \Sigma\ ,\ 1 \leqslant j \leqslant k)$叫作基于$\Sigma$ 的字符串(单词)，字符串 w 的长度用$|w|$表示，由定义可知，$|w| = k$。字符串经常用字母表中后面的字母表示(有或无索引)，如 x，y，w_j，u_i等。所有包含Σ 中字符的字符串的集合称作Σ^*。

作为一种标准的操作语言，插入操作将生成字符串的连锁或插入[57]。只有当确定的前后关系环境存在时，一个“单词”才能被插入到一字符串。更精确地说，由一系列代表前后关系环境的字符串构成一个集合，仅当一对字符串出现在一个给定的这样的集合中时，一个“单词”才能被插入。类似地，当一对字符串出现在一个这样的集合中时，删除操作可以实现将一个“单词”删去。

研究插入和删除不仅有理论上的意义，它们与实验操作也具有相关性，这也是对其进行研究的动力之一。实际上，通过合成寡核苷酸链和 PCR 定点突变技术[58]，我们可以实现在许多给定的部位进行寡核苷酸序列的插入和删除。

Kari 等[43, 51]研究了在相关位点进行插入和删除操作(下面我们简称为插入和删除)的计算学特性，研究结果表明：每台图灵机的操作都完全可以用插入和删除操作来模拟。Beaver[40]提出了一个相似的操作——碱基替换来模拟整台图灵机。

利用插入-删除系统，我们简要地描述了递归可枚举语言(等价于图灵机计算模型)的特性。这个系统根据前后关系通过插入和删除操作生成一个语言的各个元素。基于插入规则的语法已经开始从语言学角度考虑[59]。插入、删除操作对于 DNA 和 RNA 来说也是最基本的，尤其是 RNA 剪切和编辑反应[60]。我们的结果表明，这些操作，甚至是对上下关系(字符串)长度和(或)插入删除单词的长度有严格限制的操作，能完全用于计算。也就是说，它们能完全模拟任何图灵机的操作。

下面是一个插入、删除系统的构造表达式：

$$\gamma = (V,\ T,\ A,\ I,\ D)$$

这里 V 是一个字母表，$T \subseteq V$，A 是 V^*的有限子集，I 和 D 是 $V^* \times V^* \times V^*$的有限子集。

字母表 T 是γ中的最后一个字母表，A 是公理的集合，I 是插入规则的集合，D 是删除规则的集合。一个插入(删除)规则被写作一个含 3 个元素的数组$(u,\ z,\ v)$，它的意思是 z 可以在 u 和 v 之间插入或删除，这里 u 代表左侧“语境”，v 代表右侧“语境”。

对于 x，$y \in V^*$，如果下面两种情况之一发生的话，我们就说 x 源自 y，写作 $x \Rightarrow y$:

(1) $x = x_1uvx_2$，$y = x_1uzvx_2$，对于某些 x_1，$x_2 \in V^*$并且$(u,\ z,\ v) \in I$

(2) $x = x_1uvx_2$，$y = x_1uzvx_2$，对于某些 x_1，$x_2 \in V^*$并且$(u,\ z,\ v) \in D$

如果用⇒*来代表⇒关系的终止，那么由γ生成的语言可定义为：

$$L(\gamma)=\{w\in T^*|x\Rightarrow^* w,\ x\in A\}$$

非正式地说，$L(\gamma)$是通过重复应用插入和删除规则从集合 A 中得到的一系列字符串。如果：

$$\max\{|z|\,|(u,\ z,\ v)\in I\}=n,$$

$$\max\{|u|\,|(u,\ z,\ v)\in I \text{ or } (v,\ z,\ u)\in I\}=m,$$

$$\max\{|z|\,|(u,\ z,\ v)\in D\}=p,$$

$$\max\{|u|\,|(u,\ z,\ v)\in D \text{ or } (v,\ z,\ u)\in D\}=q,$$

那么我们就可以说 in/del 系统$\gamma=(V,\ T,\ A,\ I,\ D)$的权重是$(n,\ m,\ p,\ q)$。

因此，n(或 p)代表插入(或删除)序列的最大长度，而 m(或 q)代表插入(或删除)序列左侧(或右侧)“环境”序列的最大长度。

我们用 $INS_n^m DEL_p^q$，$n,\ m,\ p,\ q\geqslant 0$ 来表示由 in/del 系统[权重是$(n',\ m',\ p',\ q')$，$n'\leqslant n$，$m'\leqslant m$，$p'\leqslant p$，$q'\leqslant q$]生成的 $L(\gamma)$。如果 $n,\ m,\ p,\ q$ 其中一个参数没有边界的话，我们用∞来代替。因此 in/del 系统可以用 $INS_\infty^\infty DEL_\infty^\infty$ 来表示。

关于 in/del 所得到的主要结果如下：

theorem 1[34] $RE=INS_3^6 DEL_2^7$

theorem 2[35] $RE=INS_1^2 DEL_1^1$

theorem 3[35] $RE=INS_2^1 DEL_2^0$

theorem 4[35] $RE=INS_1^2 DEL_2^0$

定理(theorem)1 的解释是图灵机的每种操作都可以用具有有限规则的插入删除系统来模拟，这里插入字符串的最大长度是 3，插入点左右“环境”序列最大长度是 6，删除的最大长度是 2，删除点左右“环境”序列最大长度是 7。这意味着可以用 PCR 定点突变来模拟一台图灵机。定理 2~4 说明利用更小的长度(插入、删除序列的长度及其“环境序列”长度)也能得到同样的计算能力。这些结果提示我们用另一种可能的途径实现生物计算，那就是 RNA 编辑[60]，它包括了对单个核苷酸的插入和删除操作。最近的研究结果[51]表明，有限的忠实的 in/del 系统完全具有图灵机的计算能力，它只插入或删除一个字符，但其所能插入或删除的序列长度，以及环境序列的长度是没有限制的。

总之，无论是定点突变还是 RNA 编辑，它们都能成功地模拟图灵机的操作，这意味着将有许多生物计算的平台。

23.4 计算学问题所用到的自然方法

分子计算方面的研究无疑将对科学技术的许多方面产生巨大的影响，尤其是分子计算用新的方式将计算的性质清楚明白地显示出来，同时它也带来了设计一个与当今计算机完全不同的计算设备的希望。探查生物分子体内和体外计算的局限性，也许会使人们重新认识一些自然现象，如在有细胞结构的有机体中 DNA 存储信息的能力以及自然界存在的计算过程等方面。

23.4.1 RNA 编辑

我们已经提到计算过程存在于大量存在 RNA 编辑现象的单细胞或多细胞有机体中[61]。RNA 通过添加、删除或替换核苷酸来对翻译前的 mRNA 进行编辑，这种现象存在于从寄生性的原生生物到人类的许多真核生物中。例如，图 23.3 所显示的基因插入了大量的 U(RNA 由 A、C、G、U 构成，U 替换了 DNA 中的 T)。在像锥体虫这样的生物体内，RNA 编辑过程连续地添加或删除了成百个 U 残基。这些过程生成了起始或终止密码子，改变了转录的结构特征，并构造了基因编码区的 90%以上。平均来说，U 的插入或缺失占许多基因核苷酸改变的 60%。其他的核苷酸 A、C 和 G，在 DNA 和 RNA 间是完全保守的[60]。

```
DNA    G      G      GTTTTGG   AGA          G ATTTGG  A
RNA    uGu u u u GuuuuuGUUUUGGuuuAGAuuuuuuuuGuAU**GGuuAuuu
```

图 23.3 通过 u 的插入和删除进行的 RNA 编辑

将编码 *H. mariadeanei* 细胞色素氧化酶的 RNA 序列(下面的)与其基因组 DNA 序列进行比较(上面的)[60]。通过 RNA 编辑，在 mRNA 序列中加入了许多尿嘧啶，有 2 个胸腺嘧啶被删除，图中用星号代替

RNA 编辑使得编码蛋白的 mRNA 从加密的基因组片段得以恢复。引导 RNA(gRNA)这一短序列与 mRNA 前体分子间的碱基互补配对作用提供了决定插入或删除的位点(图 23.4)。令人惊讶的是，这个过程能从 12 种左右 RNA 分子产生一个保守的编码蛋白的序列，每一个包含一个特殊的环状 DNA 分子(基因本身存在于一个大环中，每一个引导 RNA 通常存在于大量小环中)。

```
                        gRNA
    3'-UaaUUaUagaaCaUagagaCUgUaaaUaaAUAAACAACCAAAUAUA-5'
       :|||||:|:||||||:|:|||::|||||||||||||||||||||:|
5'...GuuAAuGuuuuGuAuuuuuGAuGuuuAuuuAuuuGuuGGuuuAuGu...3'
                                                  mRNA
```

图 23.4 引导 RNA 和信使 RNA 碱基配对相互作用指导的 RNA 编辑

gRNA 序列中小写的 a 和 g 通过碱基配对引导 mRNA 序列中插入一系列 u

对于 mRNA 中每个插入的 U，gRNA 中对应的 A 或 G 与之配对形成完全编辑的产物(图 23.4)。从 mRNA 3′到 5′的完全编辑需要一整套 gRNA。由每一个 gRNA 引导的编辑导致一系列重叠的插入和删除，实际上构成了基于 RNA 的计算机[61]。

23.4.2 基因的翻译

纤毛类型的原生动物有两种类型的核：一个有活性的大核和一个无活性的小核(只负责繁殖)，大核在繁殖后从小核发育而来。无纤毛原生动物小核中编码蛋白的基因由于非编码蛋白 DNA 序列(中间去除序列)的插入而被掩盖，这些序列必须在组装成有功能的大核 DNA 前除去。尖毛虫(*Oxytricha*)和棘尾虫(*Stylonychia*)中编码蛋白质的 DNA 片段(大核目的序列，MDS)有时相对于它们在大核中最后的位置发生序列的改变。例如，我们已经发现在 *S.lemnae* 中的 DNA 聚合酶α的基因是由小核中几十个片段拼凑起来的。通过将小核中数十个片段重新拼凑起来，*O.trifallax* 解决了一个复杂的计算学问题[61]。

反拼凑的过程与 Adleman[1]用来解决“7 城”问题(定向哈密尔顿问题的一个实例)的 DNA 算法有着惊人的相似性。正在开发中的纤毛中大核“计算机”(图 23.5)利用了包含 2~14 个短的寡核苷酸正向重复序列的信息。它们引导一系列的同源重组反应。例如，MDS *n* 和下游的 IES 连接处的 DNA 序列一般与 MDS *n*+1 和其上游 IES 之间的序列是一样的，这使 MDS *n* 和 MDS *n*+1 之间正确连接。通过提供像 Adleman 图中“边”一样的“夹板”，这个机制以正确的顺序将编码蛋

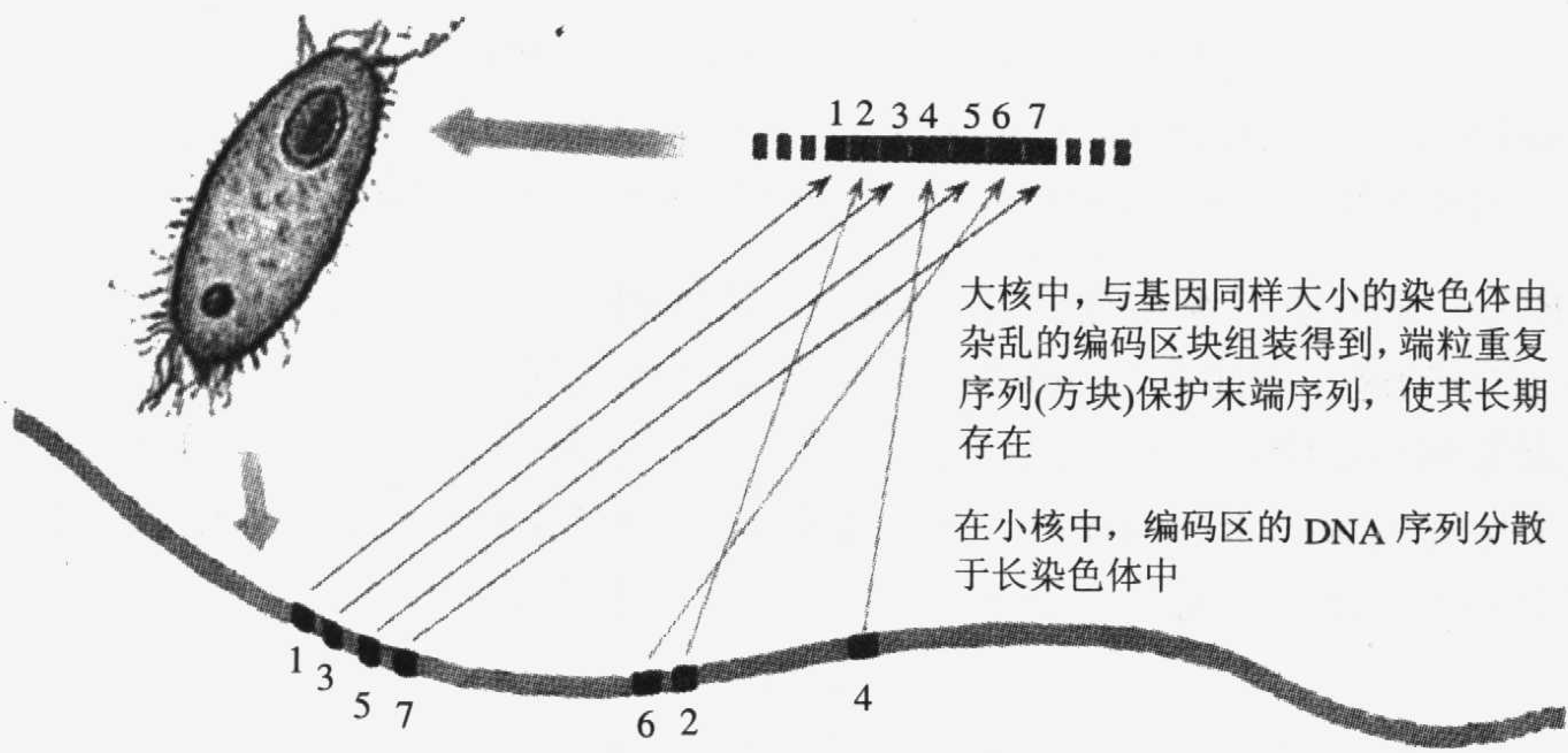


图 23.5 将基因翻译作为一个计算问题

散布的 MDS 序列，如 1-3-5-7-6-2-4(下面)，在大核形成的过程中形成了功能基因的拷贝(上面)。端粒序列的添加是基因末端的标记，并保护其末端，它相当于 Adleman 实验过程中步骤(2)的 PCR 引物，因为只有存在端粒的序列能够长期存在

白的基因组装起来，成为最终的蛋白编码序列(哈密尔顿路径)，但这其中的机制还不是很清楚。同样，细胞中基因组装可以完成惊人的计算功能，尤其是哈密尔顿路径问题，可以解决大约 50 个结点的问题，这对于其他计算机都是可望而不可及的。

RNA 编辑和基因反拼凑提供了一系列有潜在功用的生物计算方法。此外，这些过程强调存在于生物体内的计算方法的多样性，并提出了许多将生物学用于计算的模型。

(杨 冬 译)

参 考 文 献

[1] Adleman, L. (1994) Molecular computation of solutions to combinatorial problems. *Science* **266,** 1021-1024.

[la] Kari, L. (1997) DNA computing: arrival of biological mathematics. *The Mathematical Intelligencer,* **19,** 2, 9-22.

[2] Baum, E. (1995) Building an associative memory vastly larger than the brain. *Science* **268,** 583-585.

[3] Reif, J. (1995) Parallel molecular computation: models and simulations, in *Proceedings of the 7th Annual ACM Symposium on Parallel Algorithms and Architectures,* Santa Barbara, CA, pp. 213-223.

[4] Kendrew, J. (eic) (1994) *The Encyclopedia of Molecular Biology,* Blackwell Science, Oxford.

[5] Garey, M. and Johnson, D. (1979) *Computers and Intractability. A Guide to the Theory of NP-completeness.* W. H. Freeman and Company, San Francisco.

[6] Gifford, D. K. (1994) On the path to computation with DNA. *Science* **266,** 993-994.

[7] Kaplan, P., Cecchi, G., and Libchaber, A.(1995) *Molecular computation: Adleman's experiment repeated.* NEC Technical Report.

[8] Guarnieri, F., Fliss, M., and Bancroft, C. (1996) Making DNA add. *Science,* **273,** 220-223.

[9] Liu, Q., Guo, Z., Condon, A., Corn, R., Lagally, M., and Smith, L. (1999) A surface-based approach to DNA computation, in *DNA Based Computers II* (L. F. Landweber and E. B. Baum, eds.), DIMACS Series in Discrete Mathematics and Theoretical Computer Science, vol. 44, American Mathematical Society, Providence, RI, pp. 123-132.

[10] Ouyang, Q., Kaplan, P. D., Liu, S., and Libchaber, A. (1997) DNA solution of the maximal clique problem. *Science,* **278,** 446-449.

[11] Cukras, A., Faulhammer, D., Lipton, R., and Landweber, L. F. (1998). Chess games: a model for RNA-based computation, in *Proceedings of the Fourth International Meeting on DNA Based Computers* (Kari, L., Rubin, H., and Wood, D. H., eds.), University of Pennsylvania, Philadelphia, PA, pp. 27-37.

[12] Deaton, R., Murphy, R., Rose, J., Garzon, M., Franceschetti, D., and Stevens, S.(1997)A DNA based implementation of an evolutionary search for good encodings for DNA computation, in *Proceedings of the IEEE International Conference on Evolutionary Computation,* Indianapolis, IN, IEEE, Piscataway, NJ, pp.267-271.

[13] Kaplan, P., Cecchi, G., and Libchaber, A. (1999) DNA-based molecular computation: template-template interactions in PCR, in *DNA Based Computers II* (Landweber, L. F. and Baum, E. B., eds.), DIMACS Series in Discrete Mathematics and Theoretical Computer Science, vol. 44, American Mathematical Society, Providence, RI, pp. 97-104.

[14] Winfree, E., Yang, X., and Seeman, N. (1999) Universal computation via selfassembly of DNA: some theory and experiments, in *DNA Based Computers II* (Landweber, L. F. and Baum, E. B., eds.), DIMACS Series in Discrete Mathematics and Theoretical Computer Science, vol. 44, American Mathematical Society, Providence, RI, pp. 191-213.

[15] Seeman, N., Wang, H., Liu, B., Qi, J., Li, X., Yang, X., Liu, F., Sun, W., Shen, Z., Sha, R., Mao, C., Wang, Y., Zhang, S., Fu, T.-J., Du, S., Mueller, J. E., Zhang, Y., and Chen, J. (1999) The perils of polynucleotides: the experimental gap between the design and assembly of unusual DNA structures, in *DNA Based Computers II* (Landweber, L. F. and Baum, E. B., eds.), DIMACS Series in Discrete Mathematics and Theoretical Computer Science, vol. 44, American Mathematical Society, Providence, RI, pp. 215-233.

[16] Arita, M., Hagiya, M., and Suyama A., (1997) Joining and rotating data with molecules, in *Proceedings of the IEEE International Conference on Evolutionary Computation,* Institute of Electrical and Electronics Engineers (IEEE), pp. 243-248.

[17] Arita, M., Suyama, A., and Hagiya, M., (1997) A heuristic approach for Hamiltonian Path Problem with molecules, in *Genetic Programming* 1997: *Proceedings of the Second Annual Conference* (Koza, J. R., Deb, K., Dorigo, M., Fogel, D. B., Garzon, M., Iba, H., and Riolo, R. L., eds.), Stanford University, Palo Alto, CA, Morgan Kaufmann, pp. 457-462.

[18] Hagiya, M. and Arita M. (1999) Towards parallel evaluation and learning of Boolean μ-formulas with molecules, in *DNA Based Computers III* (D. H. Wood, ed.), DIMACS Series in Discrete Mathematics and Theoretical Computer Science, American Mathematical Society, Providence, RI, in press.

[19] Jonoska, N. and Karl, S. (1997) Ligation experiments in computing with DNA. *Proceedings of the IEEE International Conference on Evolutionary Computation,* Indianapolis, IN, IEEE, Piscataway, NJ, pp. 261-266.

[20] Mulawka, J., Weglenski, P., and Borsuk, P. (1998). Implementation of the Inference Engine based on Molecular Computing Technique, in press.

[21] Gloor, G., Kari, L., Gaasenbeek, M., and Yu, S. (1998) Towards a DNA solution to the Shortest Common Superstring Problem, in *Proceedings of the IEEE International Joint Symposia on Intelligence and Systems,* Rockville, MD, IEEE Computer Society Press, Los Alamitos, CA, pp. 140-145.

[22] Lipton, R. (1995) DNA solution of hard computational problems. *Science,* **268,** 542-545.

[23] Boneh, D., Dunworth, C., and Lipton, R. J. (1996). Breaking DES using a molecular computer, in *DNA Based Computers: Proceedings of a DIMACS Workshop* (Lipton, R. J. and Baum, E. B., eds.), American Mathematical Society, Providence, RI, 27, pp. 37-65.

[24] Adleman, L., Rothemund, P., Roweis, S., and Winfree, E. (1999) On applying molecular computation to the Data Encryption Standard, in *DNA Based Computers II* (Landweber, L. F. and Baum, E. B., eds.), DIMACS Series in Discrete Mathematics and Theoretical Computer Science, vol. 44, American Mathematical Society, Providence, RI, pp. 31-44.

[25] Leete, T., Schwartz, M., Williams, R., Wood, W., Salem, J., and Rubin, H. (1999) Massively parallel DNA computation: expansion of symbolic determinants, in *DNA Based Computers II* (Landweber, L. F. and Baum, E. B., eds.), DIMACS Series in Discrete Mathematics and Theoretical Computer Science, vol. 44, American Mathematical Society, Providence, RI, pp. 45-58.

[26] Oliver, J. (1997) Matrix multiplication with DNA. *Journal of Molecular Evolution,* **45,** 161-167.

[27] Baum, E. and Boneh, D. (1999) Running dynamic programming algorithms on a DNA computer, in *DNA Based Computers II* (Landweber. F. and Baum, E. B., eds.), DIMACS Series in Discrete Mathematics and Theoretical Computer Science, vol. 44, American Mathematical Society, Providence, RI, pp. 77-85.

[28] Jonoska, N. and Karl, S. (1999) A molecular computation of the road coloring problem, in *DNA Based Computers II* (Landweber, L. F. and Baum, E. B., eds.), DIMACS Series in Discrete Mathematics and Theoretical Computer Science, vol. 44, American Mathematical Society, Providence, RI, pp. 87-96.

[29] Williams, R. and Wood, D. (1999) Exascale computer algebra problems interconnect with molecular reactions and complexity theory, in *DNA Based Computers II* (Landweber, L. F. and Baum, E. B., eds.), DIMACS Series in Discrete Mathematics and Theoretical Computer Science, vol. 44, American Mathematical Society, Providence, RI, pp. 267-275.

[30] Kari, L., Gloor, G., and Yu, S. (1999)Using DNA to solve the Bounded Post Correspondence Problem. *Theoretical Computer Science,* in press.

[31] Kobayashi, S., Yokomori, T., Sampei, G., and Mizobuchi, K. (1997) DNA implementation of simple Horn clause computation, in *Proceedings of the IEEE International Conference on Evolutionary Computation*, Indianapolis, IN, IEEE, Piscataway, NJ, pp. 213-217.

[32] Adleman, L. (1996) On constructing a molecular computer, in *DNA Based Computers: Proceedings of a DIMACS Workshop* (Lipton, R. J. and Baum, E. B., eds.), American Mathematical Society, Providence, RI, pp. 1-21.

[33] Amos, M., Gibbons, A., and Hodgson, D. (1999) Error-resistant implementation of DNA computation, in *DNA Based Computers II* (Landweber, L. F. and Baum, E. B., eds.), DIMACS Series in Discrete Mathematics and Theoretical Computer Science, vol. 44, American Mathematical Society, Providence, RI, pp. 151-161.

[34] Lipton, R. (1996) Speeding up computations via molecular biology, in *DNA Based Computers* (Lipton, R. J. and Baum, E. B., eds.), American Mathematical Society, Providence, RI, pp. 67-74.

[35] Roweis, S., Winfree, E., Burgoyne, R., Chelyapov, N., Goodman, M., Rothemund, P., and Adleman, L. (1999) A sticker based architecture for DNA computation, in *DNA Based Computers II* (Landweber, L. F. and Baum, E. B., eds.), DIMACS Series in Discrete Mathematics and Theoretical Computer Science, vol. 44, American Mathematical Society, Providence, RI, pp. 1-29.

[36] Ogihara, M. and Ray, A. (1998). The minimum DNA computation model and its computational power. University of Rochester, Technical report TR-672.

[37] Beaver, D. (1995) Computing with DNA. *J. Comput. Biol.*, **2,** 1-7.

[38] Smith, W. (1996) DNA computers *in vitro* and *in vivo*, in *DNA Based Computers: Proceedings of a DIMACS Workshop* (Lipton, R. J. and Baum, E. B., eds.), American Mathematical Society, Providence, RI, DIMACS series, 27, pp. 121-185.

[39] Winfree, E. (1996) On the computational power of DNA annealing and ligation, in *DNA Based Computers: Proceedings of a DIMACS Workshop* (Lipton, R. J. and Baum, E. B., eds.), American Mathematical Society, Providence, RI, DIMACS series, vol. 27, pp. 199-221.

[40] Beaver, D. (1996) A universal molecular computer, in *DNA Based Computers: Proceedings of a DIMACS Workshop* (Lipton, R. J. and Baum, E. B., eds.), American Mathematical Society, Providence, RI, pp. 29-36.

[41] Head, T. (1987) Formal language theory and DNA: an analysis of the generative capacity of recombinant behaviors. Bulletin of Mathematical Biology, **49,** 737-759.

[42] Rothemund, P. (1996) A DNA and restriction enzyme implementation of Turing machines, in *DNA Based Computers: Proceedings of a DIMACS Workshop* (Lipton, R. J. and Baum, E. B., eds.), American Mathematical Society, Providence, RI, pp. 75-119.

[43] Kari, L., and Thierrin, G. (1996) Contextual insertions/deletions and computability. *Information and Computation,* **131,** 47-61.

[44] Yokomori, T. and Kobayashi, S. (1999) DNA-EC: a model of DNA computing based on equality checking, in *DNA Based Computers III* (in Wood, D. H., ed.), DIMACS Series in Discrete Mathematics and Theoretical Computer Science, American Mathematical Society, Providence, RI. , in press.

[45] Head, T., Paun, G., and Pixton, D. (1996) Language theory and molecular genetics, in *Handbook of Formal Languages* (Rozenberg, G. and Salomaa, A., eds.), Springer Verlag, Berlin, **2,** 295-358.

[46] Paun, G. and Salomaa, A. (1996) DNA computing based on the splicing operation. *Mathematica Japonica,* **43,** 3, 607-632.

[47] Paun, G. (1995) On the power of the splicing operation. *International Journal of Computer Mathematics,* 59, 27-35.

[48] Freund, R., Kari, L., and Paun, G. (1999) DNA computing based on splicing: the existence of universal computers. *Theory of Computing Systems* **32,** 69-112.

[49] Csuhaj-Varju, E., Freund, R., Kari, L., and Paun, G. (1996). DNA computing based on splicing: universality results, in *Proceedings of 1st Annual Pacific Symposium on Biocomputing,* Hawaii (Hunter, L. and Klein, T., eds.), World Scientific Publ., Singapore, pp. 179-190.

[50] Yokomori, T., Kobayashi, S., and Ferretti, C. (1997)On the power of circular splicing systems and DNA computab-

ility, in *Proceedings of the IEEE International Conference on Evolutionary Computation,* Indianapolis, IN, IEEE, pp.219-224.

[51] Kari, L., Paun, G., Thierrin, G., and Yu, S. (1999) At the crossroads of DNA computing and formal languages: characterizing recursively enumerable languages using insertion/deletion systems, in *DNA Based Computers III* (Wood, D. H., ed.), DIMACS Series in Discrete Mathematics and Theoretical Computer Science, American Mathematical Society, Providence, RI, in press.

[52] Hartmanis, J. (1995) On the weight of computations. *Bulletin European Association of Theoretical Computer Science,* **55,** 136-138.

[53] Kurtz, S., Mahaney, S., Royer, J., and Simon, J. (1999) Active transport in biological computing, in *DNA Based Computers II* (Landweber, L. F. and Baum, E. B., eds.), DIMACS Series in Discrete Mathematics and Theoretical Computer Science, vol. 44, American Mathematical Society, Providence, RI, pp. 171-179.

[54] Amenyo, J. (1999) Mesoscopic computer engineering: automating DNA-based molecular computing via traditional practices of parallel computer architecture design, in *DNA Based Computers II* (Landweber, L. F. and Baum, E. B., eds.), DIMACS Series in Discrete Mathematics and Theoretical Computer Science, vol. 44, American Mathematical Society, Providence, RI, pp. 133-150.

[55] Mihalache, V. (1997) Prolog approach to DNA computing, in *Proceedings of the IEEE International Conference on Evolutionary Computation,* Indianapolis, IN, IEEE, pp. 249-254.

[56] Salomaa, A. (1973) *Formal Languages.* Academic Press, New York.

[57] Kari, L. (1991) *On insertions and deletions in formal languages.* Ph.D. thesis, University of Turku, Finland.

[58] Dieffenbach, C. W. and Dveksler, G. S., (eds.), (1995) *PCR primer: a laboratory manual,* Cold Spring Harbor, NY, Cold Spring Harbor Laboratory Press, pp.581-621.

[59] Galiukschov, B. S. (1981) Semicontextual grammars (in Russian). *Mat. logica i mat. ling.,* Kalinin Univ., 38-50.

[60] Landweber, L. F. and Gilbert, W. (1993). RNA editing as a source of genetic variation. *Nature* **363,** 179-182.

[61] Landweber, L. F. and Kari, L. (1998) The Evolution of DNA Computing: Nature's Solution to a Combinatorial Problem, in *Genetic Programming* 1998: *Proceedings of the Third Annual Conference, July* 22-25, 1998, (Koza, J. R., Banzhaf, W., Chellapilla, K., Deb, K., Dorigo, M., Fogel, D. B., Garzon, M. H., Goldberg, D. E., Iba, H., and Riolo, R. L., eds), University of Wisconsin, Madison, WI, San Francisco, CA, Morgan Kaufmann, pp. 700-708.

24 检测生物模式:整合数据库、模型和算法

Gautam B. Singh

24.1 引言

生命体的每个细胞基本上包含着相同的基因组。然而，某一种细胞中所表达的基因不同于其他种类的细胞，这是由细胞所行使的功能决定的。我们也知道，真核生物基因组中仅有 10%~20%的 DNA 编码蛋白。研究表明大部分 DNA 是不编码的，但对于基因表达调控是非常重要的。而且，基因调控的机制是与具有生物意义的 DNA 序列模式相配合的，这些序列模式存在于基因间的非编码区[1]。

在 DNA 的非编码区发现了许多重要的特殊调控序列，如内含子、启动子、增强子、基质结合区(MAR)以及重复序列等。这些区域中许多包含着一些 DNA 模式，它们体现细胞特异基因表达功能控制点[2, 3]，而其他的 DNA 模式，如重复序列可能作为一个生物钟[4]。这些例子以及大量的其他实例表明真核生物 DNA 中的模式序列扮演着十分重要的角色。其他 DNA 模式的例子包括富含 AT 或 GC 的区域、端粒重复序列 AGGGTT(人 DNA 内)、4 核苷酸 CTAG 以及基因编码区的 GNN 周期。这些证据表明模式序列的变化对于机体的活性是有害的。

因此，DNA 并不是均匀的字符串，而是镶嵌着许多序列水平的基序，它们相互作用，调节蛋白的合成。下面的 4 个事件是以 DNA 为模板转录为 RNA 所必需的：

(1) 基因的增强或 DNA 结构的识别：这个步骤是进行表达的先决条件。实际上，被增强了的基因座落在 10nm 的纤维上，能够被转录。

(2) 转录的起始，即 RNA 聚合酶与 DNA 双链结合：这一步在 RNA 聚合酶与 DNA 结合的区域(称为启动子)，DNA 解为单链。

(3) 延伸，即在不断延伸的核苷酸链的 3′ 端共价连接新的核苷酸。

(4) 终止，识别转录终止序列并释放 RNA 聚合酶。

DNA 的结构特性是维持细胞内一个基因增强的主要原因。尽管 RNA 聚合酶是转录的真正执行者，但其他被称为转录因子的蛋白对于这个过程的起始也是必需的。这些因子有的直接辅助 RNA 聚合酶，有的在转录元件的组装过程中起辅

助作用。这就是说，转录因子通过与 RNA 聚合酶结合、与其他转录因子结合以及与 DNA 顺式序列结合等方式促进转录的进行。

转录因子进一步分为三类：基本的、上游的、可诱导的。基本转录因子是所有的基因都需要的，对于转录过程来说是必需的。TATA 框和 CAAT 框就属于基本转录因子。上游转录因子影响转录的效率，具有这种因子的细胞核能大量地合成基因。可诱导的转录因子可被许多刺激物激活或在生长期被激活以调节基因的表达。

原核生物和真核生物编码蛋白的基因(如通过 RNA 聚合酶 II 转录的基因)转录所需的基本转录装置具有相似性，但它们在两种细胞中的位置不同。一般情况下，原核生物启动子处于开放状态，对其控制属于负调控方式。相反，一个典型的真核生物启动子虽然能启动基本的转录，但需要正调控因子提高其转录效率。这种上游结合的转录因子一般会增强转录水平，并且对于一个启动子经常是特异的。它们通过与基本的转录装置相互作用并与辅助激活因子结合来辅助转录的起始。图 24.1 给出了一些基本转录因子和上游转录因子的 DNA 结合位点。

A

细胞类型	位置	序列
原核生物	~-10 bp	TATAAT
	~-35 bp	TTGACA
真核生物	~-25 bp	TATA
	~-80 bp	CAAT

基础

B

启动子名称	序列	转录因子名称
CAAT box	GGCCAATCC	CTF
GC box	GGGCGG	SP1
Octomer box	ATTTGCAT	Oct1

上游

图 24.1 启动子序列及其在 DNA 上的位点，以及基因表达所必需的转录元件形成时的辅助元件

启动子上游序列，这一类转录因子有好多种，它们的结合区很大。由于它们可以结合很大的区域，因此对其控制是很复杂的。图 24.1B 中所列的 3 种因子是普遍存在的，例如，它们存在于人体内所有的组织中。有这些因子的存在，转录的效率就大大提高了。有时这些位点甚至对于体内转录是必需的。另外，一些较远的影响转录的元件称作增强子。与上游启动子元件不同，增强子在基因周围的任何位置结合都能促进转录。

最后一种转录因子，当受到环境因素刺激时被激活，因此是基因表达的最后一个控制点。这些因子称为诱导因子。真核生物启动子除了具有组成性的转录因子 TFIID、SP1 等结合位点外，还具有诱导因子的结合位点。这些因子在受到生长因子、营养水平、热激或其他信号等因素刺激时将激活相应基因的转录。受这

些因子调控的基因在其转录起始点的 5′ 端有一些共同的结合位点。这些转录激活物通常以低聚物的形式调控表达。也就是说，一个诱导因子与其 DNA 上对应的启动子区域的结合将导致其他转录因子发挥作用，这些因子对于转录起始来说也是必需的。表 24.1 列出了一些诱导型转录因子。

表 24.1　诱导性转录因子的几个实例

刺激	作用位点	大小	转录因子名称
热激	HSE	27bp	HSF 或 HSTF
糖皮质激素	GRE	20bp	糖皮质激素受体
血清	SRE	20bp	SRF

曾经以 DNA 结合区域的结构特征对转录因子进行分类，这种分类方法产生了一系列总科(超家族)，包括 b/ZIP、bHLH、同域蛋白、ETS、REL、锌指核蛋白、HMG 家族、侧翼螺旋、血清反应因子和 CTF/NF-1 等，大部分可进一步化分成一些家族，属于特异家族的蛋白能识别特异结构的顺式元件。例如，AP-1、CREB/ATF 和 C/EBP 等家族属于 b/ZIP 超家族；EGR 和 SP-1 家族是 C_2H_2 超家族的成员。这两个超家族所识别的顺式元件具有不同的结构。转录因子的多样性对于细胞特异性分化基因调控来说是必需的。

24.2　材料

本节简要介绍了许多存储生物模式序列的数据库的情况。许多这样的数据库包含着进化信息，20 世纪 80 年代中期首次建立以来修改过多次。早期它强调模式序列的规范化，现在注重于功能信息的整合。

24.2.1　TFD：转录因子数据库

这个数据库用于管理我们在研究真核生物基因调控时所得到的序列信息。对于这些信息的管理采用了关系数据库模型，因为这种模型有助于高效地管理序列基序、模式序列结合因子、所属区域及基因/组织特异性等之间的关系。这些信息用 5 个关系表来记录：SITES、DOMAINS、FACTORS、cDNAs 和 ELEMENTS。SITES 表提供了被特异性转录因子识别的 DNA 序列的特异信息[5]。从功能的角度来考虑，认为 TFD 数据库包含两部分：第一部分包含一系列的蛋白序列，即转录调控因子的氨基酸序列；第二部分是一个模式序列数据库，存放被转录因子识别的核苷酸模式序列。TFD 的内容通过一些 DNA 分析软件(如 GCG、Milwaukee、WI)可以得到。TFD 每年更新 4 次。TFD 与序列结合位点的

入口提供了下列信息[6]:

UAS(G)-pMH100　CGGAGTACTGTCCTCCG　GAL4　J. Mol. Biol. 209, 423~432(1989)

TFIIIC-Xls-5S.1　TGGATGGGAG　TFIIIC EMBO. J 6,3057~3063(1987)

GCN4-his3-189　ATGACTCAT　GCN4　Science 234,451-457(1986)

ooTFD(面向对象转录因子数据库)是 TFD 的最新继承者，它应用面向对象技术来描述转录因子、相关蛋白、cDNA 以及相关文献之间的关系[7]。特别是，ooTFD 应用这项新技术描述了容量和成分——绘制相关计划时一些麻烦的关系。通过这种方式，ooTFD 能够描述出所有转录因子的信息(包括真核生物、原核生物、基本的以及由多蛋白复合物或单体蛋白构成的调节因子)。ooTFD 及其相关工具可以从 http://www.isbi.net 得到。

24.2.2 TRANSFAC

开发 TRANSFAC 数据库的主要目的是为理解基因组序列中发现的调控信号的功能提供相关的生物学背景信息。应用这些信号是为了提供所有调控蛋白的相关数据，并使研究者可以逐级追溯转录控制的由来[8, 9]。TRANSFAC 数据库包含 DNA 调控序列，以及与之结合并发生作用的转录因子的信息。这个数据库用于描述上述元件，还用于为特定功能元件定义共有序列和矩阵，并提供了在一个未知的基因组中鉴定调控信号的方法[10~12]。

当前，由 TRANSFAC 发展出了3 个数据库，它们分别是：TRANSFAC、TRRD(转录调控区域数据库)和 COMPEL(复合元件数据库)。TRANSFAC 是一个关于转录因子及其 DNA 结合位点的数据库，并以关系型图表存储。根据 6 个文件得到这些关系，它们分别是 SITE、FACTOR、CLASS、MATRIX、CELL 和 GENE，它们在语义学上与 TFD 数据库非常相似。附加的 MATRIX 文件提供了一个关于结合位点信号的图谱，这对于信号搜索过程具有理论意义。SITES 和 FACTORS 两个文件中包含着指向外部数据库的链接，外部数据库包括：EPD、SwissProt、EMBL 和 PIR。TRRD 分级提供了真核生物转录调控区域结构信息以及基因表达序列特殊模式的信息，还提供了共同协作发挥作用的因子及相隔几百个碱基的启动子和增强子组织形式的信息。COMPEL 是一个关于脊椎动物基因调控元件的数据库，这些元件位于一个基因的转录控制区，且包括两个位置接近的不同转录因子结合位点。这些数据库可以从以下网址得到：http://transfac.gbf.de/TRANSFAC,http://www.bionet.nsc.ru/TRRD,http://www.bionet.nsc.ru/COMPEL[13]。

24.2.3 PROSITE 和 EPD(真核启动子数据库)

PROSITE 是一个在蛋白序列中发现功能位点和类型的复杂数据库，其开发动机源于对未知功能蛋白(来源于对基因组序列及 cDNA 序列的翻译)的功能鉴定[14, 15]。这个数据库包括具有生物学意义的模式和图谱，它们能使人建立蛋白家族，它可能包括未知的序列和未知的结构域[16]。

EPD 数据库是一个有注释的非冗余性数据库，其内容是实验中得到的真核生物 RNA 聚合酶 II 对应的启动子。EPD 中所有的信息都基于生物学实验中所论述的实验事实。EPD 通过对 EMBL 中侧翼序列、启动子鉴定的根据、其他数据库及生物学文献等加以限定来对 EPD 数据库内容进行注释。EPD 最初是为进行序列比较分析而设计的。结果，EPD 加快了具有生物意义启动子单元的快速地动态增加。这个数据库可通过下面的网址得到：`http://cmpteam4.unil.ch`[17]。

24.2.4 MAR 数据库

基质或骨架结合区是相关的较短的序列(100~1000bp)，这些序列将染色质环锚定于核基质。MAR 包括复制起始区(ORI)，并拥有集中的转录因子结合位点[18]。哺乳动物细胞核中大约含有 100 000 个基质结合位点，其中大约 30 000~40 000 是 ORI 序列[19]。已观察到 MAR 通过侧面与基因末端区域相连接，包括许多转录单元。MAR 集聚了染色质的转录活性区域，因此转录起始于核基质表面一致的区域[19, 20]。

研究者们希望像 MAR 这样的指示区域能在注重转录图谱绘制的后基因组计划中发挥重要作用。由于 MAR 在遗传过程及其在染色体功能区域定位中发挥重要作用，因此建立一种方法来将这些标记定位在测序得到的图谱上是非常有意义的。当前没有已知的与 MAR 一致的序列，但已经用实验的方法确定了一些 MAR 的基因位点，它们包括鸡溶菌酶基因[21]、人β干扰素基因[22]、人β球蛋白基因[23]、鸡α球蛋白基因[24]、p53 蛋白基因[25]以及人的精蛋白基因簇[26]。研究者们发现了作为 MAR 特征标志的基序，并对它们进行了广泛的研究。这些基序按功能进行了分类，并以 AND-OR(与-或)模式描述如下。

AND-OR 模式用逻辑或、逻辑与对序列中检测到的基序进行规范化。序列水平的基序是用于检测更高水平模式的一种最低水平的暗示信息。通常下列操作会应用于低水平的基序：

- 一致的基序序列，用 m 表示，它是正常的表达式。
- 两基序 m_i 和 m_j 的逻辑或(OR)关系用 $m_i \vee m_j$ 表示。
- 两基序 m_i 和 m_j 的逻辑与(AND)关系用 $m_i \wedge_a^b m_j$ 表示。参数 a 和 b 用于确

定两个基序间可以分离的程度。

- 基序 m 的逻辑非，用 $\bar{m}$ 表示。表示缺少一个给定的基序。

在这样的一个通用框架下，模式描述语言定义为这样的语言：即便随着我们更好地理解 DNA-蛋白质相互作用以及遗传机制控制，它也有能力描述许多易于发现的模式，与每个基序(模式)还相关的是其随机出现的概率，这个值可由被分析序列的碱基组成得到[27]。应用 AND-OR 方法，得出了下列 MAR 规则。

(1) 复制起始序列(ORI)规则：复制与核基质相关，复制起始区共有 ATTA、ATTTA 及 ATTTTA 等基序。

(2) 弯曲 DNA：已鉴定弯曲 DNA 在基质附件点或其附近，参与 DNA-蛋白质的相互作用，如重组、复制、转录等[18, 28]。已预测出具有重复基序($AAAAn_7AAAn_7AAAA$ 和 TTTAA)的序列的最适弯曲度。

(3) 纽结 DNA：纽结 DNA 的特征是存在大量的二核苷酸 TG、CA 或 TA，它们被 2~4 或 9~12 个其他核苷酸间隔开。例如，利用基序 TAn_3TGn_3CA(TA，TC，CA 出现的顺序任意)可以识别出纽结 DNA。

(4) 拓扑异构酶 II 作用位点：研究表明，拓扑异构酶 II 的结合及剪切位点同样也出现于核连接部位。脊椎动物和果蝇的拓扑异构酶 II 共同的基序也用于发现基质结合区[29, 30]。

(5) 富含 AT 的序列：对于许多 MAR 来说，富含 AT 是其典型的特征，而且富含 AT 的序列必须周期性的出现。

(6) 富含 TG 的序列：一些富含 TG 的片段也具有指示 MAR 的作用，这些区域大量存在于许多基因的 3′端非编码区(3′ UTR)，而且可能是重组位点的信号[18]。

24.2.4.1 模式的规范

接下来我们将讨论把基序的可变性比作模式来体现的问题。这个可变性可以用前面说过的 AND-OR 规则来描述。下面是一个例子，我们将这个规则用于定义 DNA 复制区。可用一个基于 OR(∨)操作符的表达式表示 3 个基序：m_1 = ATTA，m_2 = ATTTA，m_3 = ATTTTA。这个例子中没有用到 AND 操作。

$$R_1 = m_1 \vee m_2 \vee m_3 \tag{1}$$

相似地，出现多个基序时，也可用 AND(∧)来定义。利用 AND 操作，引入了另一个参数来限制 2 个共同出现的基序之间所允许的间隔。例如，富含 AT 的区域可以用 2 个同时出现的 6 核苷酸字符串来表示，m_4 = WWWWWW[注意，国际理论和应用化学联合会(IUPAC)规定用 W 代表 A 或 T]，它们之间相隔 8~12 个核苷酸，可以用 AND 操作符表示两个基序之间的距离：

$$R_2 = m_4 \overset{12}{\underset{8}{\wedge}} m_4 \tag{2}$$

DNA 序列中出现基序的意义与其纯粹偶然出现的可能性是不相关的。通过一个规则，将潜在模式随机出现的概率也用于数学统计来估计一个模式随机出现的频率。举个简单的例子，上面两个规则所描述的模式可以用于计算。公式 R_2 所描述的模式中的参数值由两个基序间所允许的距离决定。这样，上面 R_1、R_2 的随机概率分别是：

$$Pr(R_1) = Pr(m_1) + Pr(m_2) + Pr(m_3)$$

$$Pr(R_2) = Pr(m_4) \cdot \{1 - \exp[-5 \cdot Pr(m_4)]\} \tag{3}$$

相似地，由潜在基序建立的随机概率公式可以用生成的函数来计算[27]。正如下节所讲，概率公式用于估计 DNA 序列中出现的模式的统计意义。我们以用于检测 MAR 的高级模式数据库进行了描述。

24.3 方法

24.3.1 模式检测软件

“信号搜索数据”算法的产生提供了一个通用的方法，用于描述可能有相似功能的一系列 DNA 序列的特性[32]。除了对这个方法进行了详细地描述，还提供了两个应用这个算法的程序，当提供输入数据时，第一个程序将这些数据转换成有限图谱，另一个程序列出了可能有潜在功能的信号区域。

24.3.2 模式检测的网络工具

(1) Signal Scan(http://bimas.dcrt.nih.gov/molbio/signal)：这个网络工具由参考文献[33]中所叙述的一系列工具集合而成。这个工具可以发现被检测序列与 TRANSFAC 及 TFD 数据库中公布的序列之间的同源性。然而，这个工具不能解释信号位点的意义，所检测到的信号的价值程度可能与其长度有关。例如，可能有许多与较短序列(如 CP1)匹配的序列，因此，能得到一个较高的随机概率。而如果检测较长的信号序列，如糖皮质激素元件结合位点，将会得到较低的概率。同样，如果一个信号序列不像所报告的共同序列那样，那么这个搜索工具将会漏掉这个信号序列。

由 Signal Scan 生成的结果显示出信号序列的名称、数据库中公布的信号序列以及被分析序列中与信号序列匹配区域的第一个碱基的位置。如果知道结合因子名称的话也会显示出来，也会列出查询序列信号序列在数据库(TFD 或

TRANSFAC)中的位置。软件的输出结果可能以位置排序，也可能以与所发现位点结合的转录因子的类型排序。

(2) IMD Matrix Search(`http://bimas.dcrt.nih.gov/molbio/matrixs`)：这个软件实际上是 Signal Scan 的扩展，但转录因子结合位点利用了信息理论形成了矩阵的形式。因此，这个搜索软件将会寻找矩阵分布中碱基配对得分较高的区域[34]。每个矩阵中匹配得分高于中断分值的模式序列的第一个碱基的位置将显示出来(中断分值指为降低假阳性结果而限制的程度)。软件将报告一个 p 值，它正比于匹配序列的长度。对于同一因子的重叠位点，仅选择含有最高信息分值的位点。

(3) Tfsearch(`http://www.rwcp.or.jp/papia`)：这个程序由日本 Koyoto 大学的 Yutaka Akiyama 编写，它应用相关性计算来鉴定转录因子结合位点。本软件将具有已知转录位点图谱的 DNA 序列的相互关系应用于在查询序列中寻找信号序列。转录图谱从 TRANSFAC 数据库的 TFMATRIX 部分得到。

(4) MatInd 和 MatInspector(`http://www.gsf.de/biodv/matinspector.html`)：MatInd 工具从所得到的一系列短序列生成图谱。这样的一个图谱描述了 280 个存在于 TRANSFAC v3.4 中所描述的序列中的“入口”，并应用 MatInspector 工具分析查询序列。这个分析工具允许用户在一次运行中分析大量的序列，同时可以选择用于搜索的图谱矩阵。FastM 工具提供了这样的功能：它允许用户通过限定一定距离对转录因子结合位点进行搜索，这样可以生成 DNA 序列调控区域的模型[35]。关于本软件的详细说明参见第 18 章。

(5) MarFinder(`http://www.ncgr.org/MarFinder`)：通过定义一组由于其功能的相似性而结合在一起的模式序列，产生了对 MAR 搜索的结果。通过这样的分组，对一个给定组的搜索将在被查询的 DNA 序列中发现模式序列区域。如果在一个未鉴定的 DNA 序列中的特殊区域发现了许多功能相关模式组，我们就可以研究其功能。这个过程叫作“功能模式搜索”，利用 MarFinder 系统进行 MAR 相关模式组搜索就是对这个过程的演示[36]。

我们很自然地就会想到将模式簇的密度定义为序列跨度的性质。因此，一个滑动窗口的算法应用于度量这个值，度量结果由 W 和 δ 两个参数来确定。簇密度值在一个大小为 W 的窗口中度量，窗口中心的位置位于序列的第 x 个位置上。通过将窗口每次移动 δ 个核苷酸，来实现连续的窗口度量。如果 δ 值较小，可以应用线性法连接在每个位置(x, $x+\delta$, ..., $x+k\delta$)所生成的估算值。利用这种方式，通过一个关于 x 的函数就可以得到簇密度值的连续分布。

对每个窗口中模式簇密度的估计，在统计学上可以看作计算拒绝空假设的反概率(这里，空假设是指在一个给定的窗口中观察到模式序列的概率与一个长度为 W 且与被分析序列含有相同碱基组成的随机序列中出现模式序列的概率的差别没有显著意义)。所选用的反函数是 $\rho = -\log(\alpha)$，参数 α 是指错误地拒绝假设 H_0 的

概率。换句话说，α 代表模式簇纯粹是随机出现于窗口中的概率。将 DNA 正向和反向序列的ρ 值进行平均，即是一个给定位点的密度估计值。

为计算ρ 值，我们需要假定在一给定的序列窗口中搜索 k 种不同类型的模式序列。通常，这些模式通过公式 R_1，R_2，…，R_k 来定义。应用类似于方程(3)的概率公式，可以计算随机出现 k 种模式序列的概率。用 p_1，p_2，…，p_k 来代表这些值。下一步，我们将建立一个模式频率的随机向量 F。F 是一个 k 元代有分量的向量，$F=\{x_1, x_2, \dots, x_k\}$，这里的每一个分量 x_i 是一个随机变量，它代表具有 W 个 bp 的窗口中模式 R_i 的频率。随机的分量 x_i 假设在泊松分布中是独立的，每一个都有参数$\lambda_i = p_i \cdot W$。因此，纯粹是随机得到的频率向量 $F_{obs}=\{f_1, f_2, \dots, f_k\}$的联合概率可以这样计算：

$$P(F_{obs}) = \prod_{i=1}^{k} \left[\frac{\left(e^{-\lambda i} \lambda^{f_i}\right)}{f_i\,!} \right] \text{where } \lambda_i = p_i \cdot W \tag{4}$$

需要计算α 的步骤，以及模式序列频率大于或等于向量 F_{obs} 的积累概率(纯粹随机出现)可由方程(5)计算。这等价于对多变量进行单边积分，并且表示 H_0 被错误拒绝的概率。

$$\alpha = Pr(x_1 \geqslant f_1, x_2 \geqslant f_2, \dots, x_k \geqslant f_k)$$

$$= Pr(x_1 \geqslant f_1) \cdot Pr(x_2 \geqslant f_2) \cdot \dots \cdot Pr(x_k \geqslant f_k)$$

$$\left\{ \sum_{x_1=f_1}^{\infty} \left[\frac{(\exp^{-\lambda 1} \lambda_1^{x1})}{x_1\,!} \right] \right\} \cdot \left\{ \sum_{x_2=f_2}^{\infty} \left[\frac{(\exp^{-\lambda 2} \lambda_2^{x2})}{x_2\,!} \right] \right\} \cdot \dots \cdot \left\{ \sum_{x_k=f_k}^{\infty} \left[\frac{(\exp^{-\lambda k} \lambda_1^{xk})}{x_k\,!} \right] \right\} \tag{5}$$

$$\rho = \ln\left(\frac{1.0}{\alpha}\right) = -\ln(\alpha)$$

$$= \sum_{i=1}^{k} \lambda_i + \sum_{i=1}^{k} \ln f_i\,! - \sum_{i=1}^{k} f_i \ln \lambda_i$$

$$-\sum_{i=1}^{k} \ln \left[1 + \left(\frac{\lambda_i}{f_i+1}\right) + \left(\frac{\lambda_i^2}{(f_i+1)(f_i+2)}\right) + \dots + \left(\frac{\lambda_i^t}{(f_i+1)(f_i+2)\dots(f_i+t)}\right) \right] \tag{6}$$

方程(5)中的 p 和α 用于计算ρ 值或方程(6)中所限定的簇密度。

方程(6)中的无限和能迅速收敛，因此能适应性地进行所需精度的计算。对于较小的值λ_1，级数将缩短，这样最后一项是比一个任意小的常数ε 还小的值。图 24.2 是在分析人β球蛋白基因序列时产生的结果的一个“快照”。模式序列间的较强联合区即具有较高 MAR 潜能的区域。

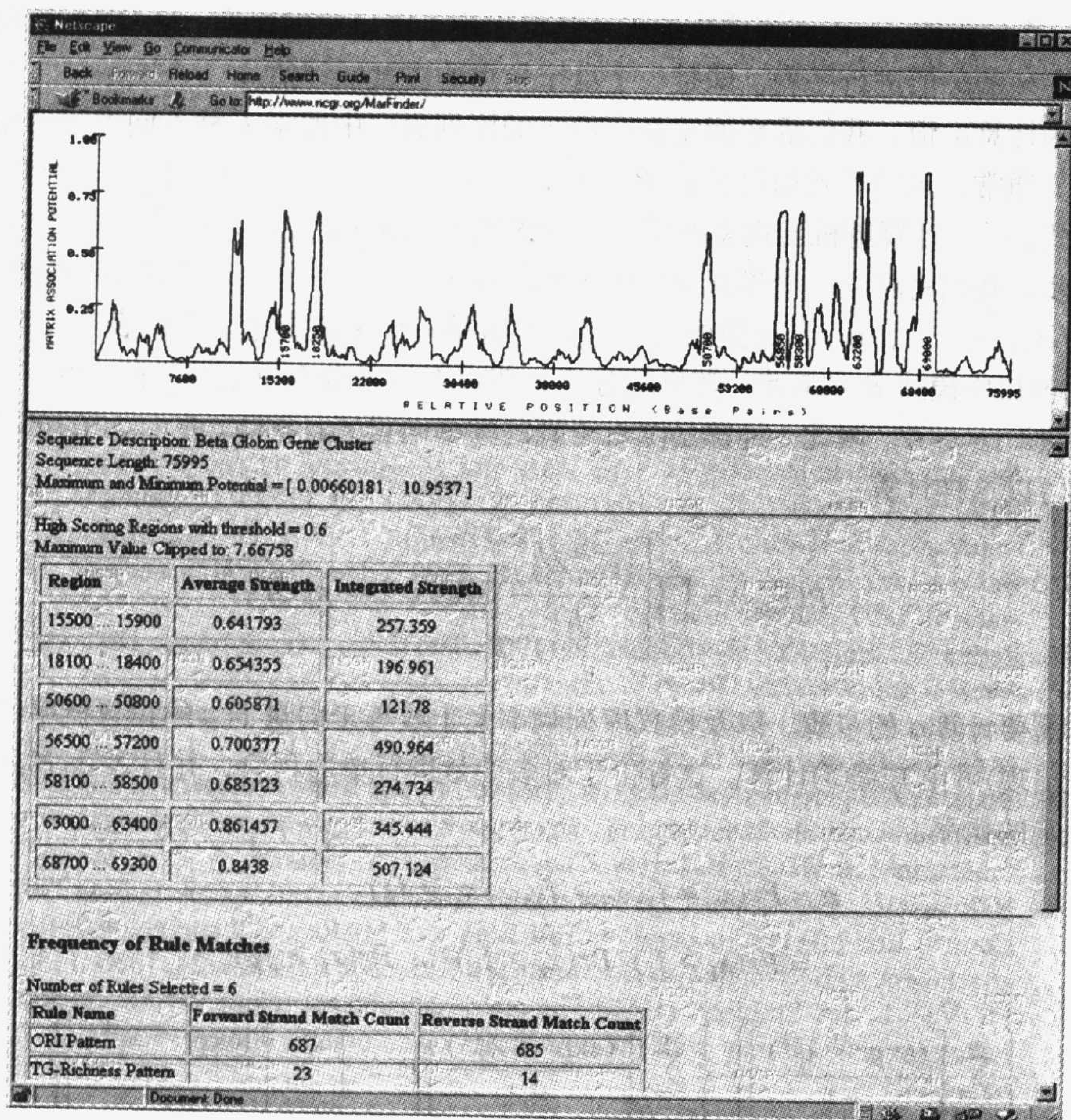


图 24.2 用 MarFinder 模式检测工具检测β球蛋白基因簇的结果。
高模式簇密度区同样也是 MAR 有意义的功能区

24.4 结论

本章总结了 DNA 序列中所出现的不同类型的序列模式。通过综合运用序列模式数据库、描述序列模式的模型，如图谱、公式等，并应用搜索的算法可以实现对这些序列模式的检测。

(杨 冬 译)

参 考 文 献

[1] Kliensmith, L. and Kish, V. (1995) *Principles of Cell and Molecular Biology,* 2nd. ed., HarperCollins, New York, NY, pp. 400-468.

[2] Kadonaga, J. (1998) Eukayotic transcription: an interlaced network of transcription factors and chromatin-modifying machines. *Cell* **92,** 307-313.

[3] Roeder, R. (1996) The role of general initiation factors in transcription by RNA polymerase II. *Trends Biochem. Sci.* **21,** 327-335.

[4] Hartwell, L. and Kasten, M. (1994) Cell cycle control and cancer. *Science* **266,** 1821-1828.

[5] Ghosh, D. (1990) A relational database of transcription factors. *Nucleic Acid Res.* **18,** 1749-1756.

[6] Ghosh, D. (1992) TFD: the transcription factors database. *Nucleic Acid Res.* **20(suppl),** 2091-2093.

[7] Ghosh, D. (1998) OOTFD (Object-Oriented Transcription Factors Database): an object-oriented successor to TFD. *Nucleic Acid Res.* **26,** 360-362.

[8] Wingender, E. (1998) Compilation of transcription regulating proteins. *Nucleic Acid Res.* **16,** 1879-1902.

[9] Wingender, E. (1990) Transcription regulating proteins and their recognition sequences. *Crit. Rev. Eukaryot. Gene Expr.* **1,** 11-48.

[10] Wingender, E. (1994) Recognition of regulatory regions in genomic sequences.*J.Biotechnol.* **35,** 273-280.

[11] Wingender, E., Karas, H., and Knuppel, R. (1996) TRANSFAC: a database on transcription factors and their DNA binding sites. *Nucleic Acid Res.* **24,** 238-241.

[12] Wingender, E., Karas, H., and Knuppel, R. (1997) TRANSFAC database as a bridge between sequence data libraries and biological function. *Pac. Symp. Biocomput.* 477-485.

[13] Wingender, E., Kel, A., Kel, O., Karas, H., Heinemeyer, T., Dietze, P., Knuppel, R., Romaschenko, A., and Kolchanov, N. (1997) TRANSFAC, TRRD and COM PEL: towards a federated database system on transcriptional regulation. *Nucleic Acid Res.* **25,** 265-268.

[14] Bairoch, A. and Bucher, P. (1994) PROSITE: recent developments. *Nucleic Acid Res.* **22,** 3583-3589.

[15] Bairoch, A., Bucher, P., and Hofmann, K. (1995) The PROSITE database, its status in 1995. *Nucleic Acid Res.* **24,** 189-196.

[16] Bairoch, A., Bucher, P., and Hofmann, K. (1997) The PROSITE database, its status in 1997. *Nucleic Acid Res.* **25,** 217-221.

[17] Perier, R., Junier, T., and Bucher, P. (1998) The Eukaryotic Promoter Database. *Nucleic Acid Res.* **26,** 353-357.

[18] Boulikas, T. (1993) Nature of DNA sequences at the attachment regions of genes to the nuclear matrix. *J. Cell. Bioch.* **52,** 14-22.

[19] Bode, J., Stengert-Ibert, M., Kay, V., Schlake, T., and Dietz-Pfeilstetter, A. (1996) Scaffold/matrix attchment regions: topological switches with multiple regulatory functions. *Crit. Rev. in Eukaryot. Gene Expr.* **6,** 115-138.

[20] Nikolaev, L., Tsevegiyn, T., Akopov, S., Ashworth, L., and Sverdlov, E. (1996) Construction of a chromosome specific library of MARs and mapping of matrix attachment regions on human chromosome 19. *Nucleic Acid Res.* **24,** 1330-1336.

[21] Phi-Van, L. and Strätling. (1988) The matrix attachment regions of the chicken lysozyme gene co-map with the boundaries of chromatin domain. *EMBO J.* **7,** 655-664.

[22] Jade, J., Rios-Ramirez, M., Mielke, C., Stengert, M., Kay, V., and Klehr-Wirth, D. (1995) Scaffold matrix attachment regions: structural properties creating transcriptionally active loci. *Intl. Rev. Cytol.* **162A,** 389-454.

[23] Jarman, A. and Higgs, D. (1988) Nuclear scaffold attachment sites in the human globin gene complexes. *EMBO J.* **7,** 3337-3344.

[24] Farache, G., Razin, S., Targa, F., and Scherrer, K. (1990) Organization of the 3'-boundary of the chicken alpha globin gene domain and characterization of a CR 1-specific protein binding site. *Nucleic Acid Res.* **18,** 401-409.

[25] Deppert, W. (1996) Binding of MAR-DNA elements by mutant p53: possible implications for oncogenic function. *J. Cell. Biochem.* **62,** 172-180.

[26] Kramer, J. A. and Krawetz, S. A. (1996) Nuclear matrix interactions within the sperm genome. *J. Biol. Chem.* **271,** 11619-11 622.

[27] Staden, R. A. (1988) Methods for calculating the probabilities of finding patterns in sequences. *Comput. Applic. Biosci.* **5,** 89-96.

[28] von Kries, J., Phi-Van, L., Diekmann, S., and Strätling, W. (1990) A non-curved chicken lysozyme 5′ matrix attachment site is 3′ followed by a strongly curved DNA sequence. *Nucleic Acid Res.* **18,** 3881-3885.

[29] Spitzner, J. and Muller, M. (1988) A consensus sequence for cleavage by vertebrate DNA topoisomerase II. *Nucleic Acid Res.* **16,** 5533-5556.

[30] Sander, M. and Hsieh, T. (1985) Drosophila topoisomerase II double stranded DNA cleavage: analysis of DNA sequence homology at the cleavage site. *Nucleic Acid Res.* **13,** 1057-1067.

[31] Singh, G. B., Kramer, J. A., and Krawetz, S. A. (1997) Mathematical model to predict regions of chromatin attachment to the nuclear matrix. *Nucleic Acid Res.* **25,** 1419-1425.

[32] Bucher, P. and Bryan, B. (1984) Signal search analysis: a new method to localize and characterize functionally important DNA sequences. *Nucleic Acid Res.* **12,** 287-305.

[33] Prestridge, D. (1991) Signal Scan——a computer program that scans DNA sequences for eukaryotic transcriptional elements. *Comput. Applic. Biosci.* **7,** 203-206.

[34] Chen, Q., Hertz, J., and Stormo, G. (1995) MATRIX SEARCH 1.0: a computer program that scand dna sequences for transcriptional elements using a database of weight matrices. *Comput. Applic. Biosci.,* **11,** 563-566.

[35] Quandt, K., Frech, K., Karas, H., Wingender, E., and Werner, T. (1995) Matind and matinspector-new fast and versatile tools for detection of consensus matches in nucleotide sequence data. *Nucleic Acid Res.* **23,** 4878-4884.

[36] Singh, G. B., Kramer, J. A., and Krawetz, S. A. (1997) Mathematical model to predict regions of chromatin attachment to the nuclear matrix. *Nucleic Acid Res.* **25,** 1419-1425.

第五部分　生物信息学教学与最新文献跟踪

25 分子生物学和遗传学的计算机应用入门课程的设计与实施

Stephen A. Krawetz

25.1 引言

计算机生物学课程正在成为许多院校必修课程体系的组成部分。显然，对计算机生物学学习感兴趣的学生在本专业知识水平上有着明显的差别。那么，如何最大限度地满足所有参与者的需要呢？此处特地为初级和高级水平的学生设计了两套不同的课程以满足不同的需要。基础水平课程是专门为那些仅仅需要获得计算机生物学的计算机技能的学生而设计的；而高级课程则是为高年级硕士生、博士生、博士后及一些专业研究者们设计的。高级课程假设这些人员已具有序列分析及计算机系统基础知识。世界各地的各种研究机构可提供数种高级课程。其中部分可在生物科学新领域网站获得，如 http://www.bioscience.org/events.htm。

本课程的特殊注意事项由冷泉港(Cold Spring Harbor)实验室提供：http://nucleus.cshl.org/meetings/。

基因组图谱计划资源中心(Genome Mapping Project Resource Centre)(U. K)：http://www.hgmp.mrc.ac.uk/About/Docs/Courses/。

共同组织的 EMB 网络课程[and the cosponsored EMBNet courses(Europe)]：http://www.icgeb.trieste.it/net/netco- urse.html。

这些高级课程在它们各自的领域中都是非常出色的。但它们并不是为了满足那些刚刚进入分子医学和遗传学领域的学生们的需求而设置的。本章则描述了为了提高这些学生的基础而设置的入门课程。

25.1.1 课程回顾和设计考虑

最初，每一个分子生物学项目对于计算机技能的需求方面兴趣的大量增加是因为快速有效的 DNA 测序技术的涌现。不久人们便意识到只有计算机才能提供解决各种数据管理和分析问题的方法。这些技能可从一位实验室成员传给另一位。然而，人们对这种技术日益增加的兴趣导致了人们对于这种知识需求的指数倍增。而且意识到提高这门课程满足低年级研究生的普遍需求比提供基础需求的个人培训更有效。为了满足这种需要，1990 年 Wayne 州大学开设了一门

名为《分子生物学计算机应用》的入门课程。该课程现已发展成为分子医学和遗传学专业在校学生所必修的核心课程，而由于对其恰当应用即可作为强有力的工具而受到必然的欣赏，而且一直以来该课程能够被很好地接受。该课程的有效性依赖于 3 个组成部分的有机结合。首先在专门的学习中心有提供便捷培训服务的能力。早期在教育培训资源中心可以由制造商资助临时租用计算机软件和硬件，这些中心只在学生学习期间提供短暂的服务。目前，这个教育培训资源中心作为医药学校图书馆系统的一部分提供这种服务。其次，每学期同时聘用一名讲师和一名教辅人员，教辅人员对这门课整体成功教学起重要作用，当在讲解和演示的过程中，当每个学生有问题需要解答时，教辅人员可在其工作界面上提供迅捷地帮助。至少每 5 个学生需要一个教辅人员。然而每 3 个学生配一个教辅人员的比例为最佳。讲师/教辅人员的概念已被证明对满足日常课程教学是成功有效的方法。因为每个讲师也可以充当教辅人员，所以可以为每个学生提供直接的相互联系机会而帮助其日常基础学习。用这种方式学院会相应地对每个学生的功底都了如指掌，进而也可以相应地调整课程教学。最后，该课程的设置是 2 周为一个周期的单独强化训练，培训者期望在此期间学生的主要精力都集中在这门课上，并要求每个学生每天至少有 3h 的演示操练，并完成布置的家庭作业，这个要求被证明是非常关键的。

25.1.2 学习目标

有效的课程设置的必需组成部分规定为以下 4 个学习目标，每个目标的达成都应使学生掌握一系列基本技能，从而提高他们在实际操作过程中利用资源的独立开发能力与自信心。学习目标如下：

(1) 介绍并加强对计算机基本操作系统、桌面环境及功能(包括文件管理、传输及电子邮件 e-mail)等操作的熟练程度。

(2) 提高访问与搜索由虚拟图书馆提供的信息资源的基本技能。

(3) 提高使用基于网页的应用程序技能来解决计算机生物学领域的一系列问题。

(4) 提高应用研究机构的核心分析系统进行基本的和其他的序列分析技能。

完成以上目标后，预期每个学员都能独立地选择适当的分析工具和/或资源来解决相关问题。

25.1.3 课程说明

目前，课程只给予一学分并且只在入门水平上教学。注册学生限制在 10 名以内，以保证一定数量的讲师与教辅人员来满足对学生有足够时间的单独教学。课程一年比一年更新反映了这一领域的进步与参加教学的教师们的兴趣。现将 1998 年的课程介绍列于下文：

“本课程将为学生提供熟悉计算机基本应用及有关概念的机会，包括分子医学中心与遗传学核心计算机应用、虚拟图书馆、各种因特网资源。学生将会接触到UNIX 操作系统、基序界面(motif)和电子邮件。接触因特网资源将要利用网络浏览器(Netscape)。内容涵盖 GCG 分析系统、虚拟图书馆、NCBI、序列和基于文本的数据库搜索、识别重复元件、多重序列的比对、具有生物学意义的序列片段的计算机预测以及程序下载与安装。本课程将演示在各种计算机环境中的电子邮件应用并为参学者提供在每个环境中发展一套有限的基本技能的机会。本课程还将演示虚拟图书馆的有效利用，包括对感兴趣的最新发表文献的检索、MEDLINE 的应用及其他期刊索引服务器，以便提供一系列相关期刊的文献，大量已获资源也将重复出现。这种强化的实验课程学习将提供大量亲手操作的训练机会。每个学习期先以概览开始，随后就是亲手操作演示，并做例题，然后期望学生们能够解决相关问题。课程评估基于每日布置的作业及本课结束时对更复杂问题的解决能力。课程评估也包括一次口头阐述和一个书面报告。注册仅仅由授课教师许可便可。”

一个简单的演示已被用来为全体教员及初学者提供一次概览，同时也会给学生们提供一个发展一些有用的基本技能的机会，以下是 2h 概览大纲：

“前言部分将为初学者熟练掌握 e-mail 概念、虚拟图书馆、因特网资源的基本应用及相关概念提供方法，每个概念都要从概览开始，随后就是亲手操作演示，并做例题。在各种计算机环境中电子邮件的应用将会被演示，并且参加者会有在这些环境中有提高一系列有限的基本技能的机会。虚拟图书馆的有效利用也将被演示。这包括对感兴趣的最新发表文献的检索、使用 MEDLINE 及其他期刊索引服务检索一系列相关期刊文献。还要总结经费获取的来源信息。也将演示一些可得到的对研究有帮助的各种因特网资源；并将讨论另一些因特网资源包括登录新闻组、NCBI 和序列与基本文本的数据库搜寻等。这种强化的实验性课程将提供大量亲手操作的训练机会。要求参学者完成以上所学知识之后就能熟练掌握网上冲浪、一些可获得工具的应用以及与同事间进行电子信息交流。班级人数限定在为提供亲手操作所能获得的计算机数量。”

参加这个概览部分的学习并不是整个课程登记报名的必要条件，但参加这段学习确实能够提高进行其他训练的热情，若有机会，会看到登记参加 2 周强化训练课程人数的增加从而证实了这一点。

25.1.4 课程大纲和作业示例

课程大纲见表 25.1。该表提供了 1998 级每一节课的教学大纲(纲要)，该大纲

为各班提供课程计划核心内容的指导。这些课程内容包括计算机基础、重要的分子遗传学互联网站点的点评、相似序列的搜索、多重序列的比对、虚拟图书馆、高阶序列分析的搜索，以及 GCG 分析系统入门。之后，为最佳限度地满足学生群体，对本课程进行因材施教，讲师可以采用这本书的任意章节作为额外补充教材教学。每部分课程从互联网站点获得的数据分析的程序见表 25.2 中加星号的条目，每节课的问题举例见表 25.3，这些示例也可以用来做指导。如表 25.4 所示(格式部分)每节课留出 1 个多小时的时间用于答疑及对需要帮助的学生进行单独辅导，这种做法被证明是保证学生理解所讲授的内容的有效方法。

表 25.1　课程安排

第一部分：计算机基础
课程介绍与欢迎
MS　Windows 基础
UNIX　基础
PC　X-Windows(NCD　PCXWARE)
文件管理
e-mail(PINE，EUDORA，NETSCAPE MAIL)
文件传输(ftp，WSftp)
第二部分：重要的分子遗传学互联网站点总结、相似序列搜索、多重序列的比对
因特网全球网络(Word Wide Web)资源(列表及说明见表 25.2)
相似性搜索用 BLAST/FASTA
下载与程序安装(Tree Tool)
多重序列(排列)比对(CLUSTAL W 和 GeneBee)
第三部分：虚拟图书馆
搜索国家生物技术信息中心的 PubMed 系统的 MEDLINE 获取生物技术信息
搜索科学信息研究所的 Science Citation Index(SCI)和 Current Contents Connect
用文献数据库和目录服务平台锁定当前生物医学文献
获取因特网上的全文期刊及打印文献
在因特网上查询资助和基金资源
第四部分：高阶序列分析搜索简单重复序列限制性位点分析
MARFinder
识别重复元件
标识转录因子可能的结合位点
第五及第六部分：GCG 序列分析或其他可比性程序
GCG 介绍：序列分析
GCG 手册：http：//cmmg.biosci.wayne.edu/gcg/gcgmanual.html
SeqLab: GCG 的 X 界面
SeqWeb: GCG 的 Web 界面
基本序列分析
多重序列分析
第七~第九部分：最后作业安排
第十部分：论文与总结报告

表 25.2 因特网上一些有用的站点

数据库与搜索工具

NCBI

*http://www.ncbi.nlm.nih.gov/

EMBL SERVER(服务器)

*http://www2.ebi.ac.uk/services.html

基因组导航器：啤酒酵母基因组索引

http://www.mpimg-berlin-dahlem.mpg.de/~andy/GN/S.cerevisiae/

序列比对

GENEBEE 多重序列比对

*http://www.genebee.msu.su/

TREEVIEW

*http://taxonomy.zoology.gla.ac.uk/rod/treeview.html

CLUSTAL W

*http://www.2.ebi.ac.uk/clustalw/

GENEDOC:多序列比较编辑，Windows 的分析及阴影工具

http://www.cris.com/~ketchup/genedoc.shtml

序列分析

限制酶切位点消化

Webcutter2.0: 对序列进行分析并且直接援引 REBASE

*http://www.firstmarket.com/cutter/cut2.html

用 MatInspector v2.2 搜索可能的转录因子结合位点：

*http://transfac.gbf-braunschweig.de/

MAR-Finder:用 MAR-Finder 推导 DNA 序列中存在的基质相关区域或 MARs

*http://www.ncgr.org/MarFinder/

Sanger 中心信息部计算基因组小组(Computational Genomics Group of the Sanger Centre Informatics Division)

http://genomic.sanger.ac/uk/

BCM 搜索启动程序

*http://kiwi.bcm.tmc.edu:8088/search-launcher/launcher.html

重复元件

RepeatMasker2 网页服务器

*http://ccr-081.mit.edu/Repeats.html

CENSOR 网页服务器

*http://charon.girinst.org/~server/censor.html

图像分析，实验手则及计算机课程

NIH 的 IMAGE 程序(MAC & PC 版本都有)(免费的)

http://www.scioncorp.com/

PCR 及多重 PCR：指南与问题解答

http://info.med.yale.edu/genetics/ward/tavi/PCR.html

续表

欢迎到 VSNS 的 BioComputing 部门 `http://www.techfak.uni-bielefeld.de/bcd/welcome.html`
其他有用的网站
真正十分有用的 MolBioPage `http://www.lars.bbsrc.ac.uk/plantsci/molbiol/molbiol.html`
Alex's Cyber-Science Jumpstation `http://www.flnet.nl/~bossers//`
在线分析工具 `http://www-biol.univ-mrs.fr/english/logligne.html`
欢迎使用 Globin 基因服务器 `http://globin.cse.psu.edu/`
ExPASy 分子生物学服务器:瑞士生物信息研究所(SIB)2-D PAGE `http://expasy.hcuge.ch/`

* 用于数据分析的因特网站点

表 25.3 作业

作业 1:

a. 给班上的每一位同学包括教员发一条 e-mail 信息，做自我介绍并介绍你的研究项目以及为何学习本课程

b. 注册成为 `http://www.ncgr.org/MarFinder/`的一个用户

c. 画图表示 GenBank 序列文件并用注释说明

作业 2:

a. 从因特网的其中之一站点上恢复并安装一个序列编辑器，并准备在课堂上演示其功能

b. 在基因簇 U15422 中有 3 个基因，用至少 2 个不同的基于万维网资源确定本基因簇中最佳序列比对，这些比对中是否是最佳、你能进一步改进比对结果吗?

作业 3:

检索已知有关核基质序列的全部文献中的文章，从中检索相关序列

作业 4:

a. 在 U15422 序列中鉴定可能的转录因子结合位点，它们在该区域结构中的相互关系如何，哪一个位点为合理的生物学活性位点，为什么?

b. 比较和对比在 U15422 中所包含的重复元件，用至少 2 个不同的基于万维网资源鉴定之

作业 5:

选择你所感兴趣的编码蛋白的 DNA 序列，比如在实验室你正在研究的基因或其他任何基因。请准备 DNA 序列的限制性酶切位点图谱，然后在 NCBI 上进行数据库搜索，从而搜寻到与你的基因表达蛋白产物相关的其他蛋白，将你的分析结果在下次课之前通过 e-mail 发送出去，请包括对你选择的 DNA 序列及你选择的原因的说明，说明获得你的分析的具体过程，并解释结果

作业 6:

用 GCG 多序列分析工具比对你感兴趣的一组相关蛋白的氨基酸序列，从而比较它们序列的相似性和系统发生的相关性。储存树状图(dendogram)和种系图(phylogram)并将这些文件输入你的实验室或图书馆的计算机，并用 PostScript 和惠普图印打印机打印出几份以备下次课使用

最后作业:

给出人第 16 号染色体的序列表，在这些序列中，鉴定出与核基质相关的合理的可能区域以及这些位点间的相互关系

表 25.4 每日学习计划

0.5 h——提问及个别辅导
1.0 h——指导
0.5 h——休息/提问及个别辅导
1.0 h——指导
课后——提问及个别辅导

25.1.5 授课大纲

讲解从程序的背景理论开始，接着是学生们跟着讲师的演示在其各自的工作站上亲手操作演示，程序执行之后的分析结果可以作为课堂教学部分而加以阐释。若时间允许，可以调整程序参数以便说明其对输出及结果阐释的作用。

虽然学生们应当熟悉基因结构的基本概念，但他们常常并不熟悉序列元件沿着核苷酸序列链的顺序排列也可以定义一个功能性基因的概念。通常提供对序列结构、信号肽、基序修饰基因定位以及这些元件是如何进行手工鉴定的简要总结非常有帮助。这可以从核酸数据库获得的序列文件中文件头(header)信息处画线注释来做到。此外，更重要的是，必须强调的计算机输出的仅仅是生物学的预测，输出结果仅仅用于指导实验，而预测则要求进行生物学核实。

原则上各种方法与策略可用于解决本课程的问题，例如，当讨论序列相似性搜索时，会同时拿出 BLAST 和 FASTA 程序[1, 2]来讨论各种可选择程序，各种搜索策略的不同及对分析的影响也给予讨论，如何调整程序参数，如字节的大小、窗口大小和阈值，可以也予以讨论，以尽可能完善输出结果以解决问题。计分矩阵的概念，即一个矩阵是如何产生及怎样使用的，在各种程序，如 BLASTP(蛋白质对蛋白质数据库)、BLATN(核酸对核酸数据库)及 BLASTX(6 核苷读框的翻译对蛋白质数据库)中都有体现。搜索结果及相似数量的阐释也予以讨论。此外，给学生显示简单的统计学概念如 E,期望值(在随机出现的图案或事件中期望出现次数)也可以作为最终产生比对结构的一个量化指标。当演示数据库搜索策略时，对于一些远缘成员的生物学相关序列列表回顾是很有价值的。要强调的是，要小心地阐释结果，并且这必须在将问题输入计算机时搞清楚。

如上所述，数据的阐释在全课程要经常强调。最好的例子之一是搜索转录因子结合位点[3]。在这种情况下，学生总是要面对大量的必须手工分类整理的生物学上可能的位点，这种练习可以使那些知道基本的基因结构的学生开窍，如揭示与起始密码子相关的起动子的位置。这些学生可用这一信息来限定这些生物学相关因子的搜索区域。此外，对于组织分布和已经识别的因子的特异性等基本的生物学概念，可以通过提以下简单问题来予以强调。可能的转录因子有和基因研究一样的相似组织分布吗？

25.1.6 其他建议的论题

有 3 个因素在这门课程中没有论及，它们是序列编辑器、序列项目管理及系统发育分析的应用。

首先，在演示文件格式结构和文件格式转换时，期望学生在需要时能熟悉序列编辑器。绝大多数学生最初是用文本编辑器自行解决这一问题的，然而，一旦他们开始做更为复杂的作业时，序列编辑器的必要性就变得更加明显了。以我们的经验得知即使序列编辑器是必要的，大多数学生对这一节课的学习如同第一节课缺乏许多基础知识一样担忧，若这种技能相当必要，那么，就可先以文本编辑器开始，用文字处理器来加强各种程序的强制性格式要求。然后他们就能应用序列文件编辑器进行更复杂的编辑功能。在 GGG 系统(第 1、2 章)和 Staden Suite[4, 5](第 7 章)里可以找到具体的完整功能编辑器的范例。此外，如 Sequin[6]编辑器用于标准数据库呈递可能也是合适的，后者会使学生更熟悉包含在 GenBank 文件头里的大量信息资源。

其次，序列项目管理提出了明显的挑战，其唯一性通常得不到人们的理解，直到个人开始从事测序项目。因此，就要求把序列计划管理作为独立的课程。然而，一旦大多数学生熟悉了基础知识，就能很容易地理解像 Staden Suite(第 7 章)这样的序列管理软件。

第三，用 CLUSTAL W 和 tree tool 进行系统发育分析的实践操作方面已讨论过，而本书第 12 章中给出的综合性分析本章中并没有涉及。一旦学生们掌握了基础知识后，此论题将作为一门整体课程来进行讨论。

25.1.7 用帮助指南制作说明文件

绝大多数说明书是根据 Web 站点程序中便捷可用的在帮助文件的协助下准备出来的。这些程序文件为程序操作提供了非常好的资源，它们常常详细说明了程序分析、输入与输出参数、数据说明以及文件格式问题的使用方法。在适当的时候还可以从当前文献中收集额外信息。有趣的是，过去的经验表明，学生喜欢在课程开始之前就与布置的读物清单一起收到一个打包的说明书，而不愿意从 Web 网上检索同样的文件。

25.1.8 课程评估

与教辅人员的每日接触及每个学生通过 e-mail 呈交的每日作业可用于评价学生的进步，并了解需要纠正的地方。同时，学生提供第三方不署名的针对每堂课的及时反馈也是很有帮助的，这种方法可同时向讲师和教辅人员展示出结果，从学生的视点，这门课进行的怎样，以及依照他们自己的观点是如何理解所提供的材料的。这三项标准能用来指导对以后课时的教学做适当的调整或重复讲授多数

学生还未搞清的概念。每一讲之后典型的调查表见表 25.5。

表 25.5 每日课程评估

讲师：　　　　　　日期：

请完成下列评估，选择你对每个问题的反应

1. 对这次讲课你准备了吗?

　　是　　　　　否

2. 你读过这次课所有指定的有关材料吗?

　　是　　　　　否

3. 你发现该课程所覆盖的材料有用吗?

　　是　　　　　否

4. 表达清楚吗?

　　是　　　　　否

5. 是否分配了足够的时间来讨论?

　　是　　　　　否

6. 你是否理解本书/章/课所覆盖的这些材料?

　　是　　　　　否

7. 如果你对这些问题中的任何一个回答是“否”的话，请具体说明!

8. 请注明本次课的讲师:

9. 请注明本次课的教辅人员:

25.1.9 结论

本课程为计算机分子生物学领域提供了一个概览，而不是一个终结。本课程的作用在于使学生熟悉许多可利用的资源，并逐步培养这样一个概念：很多不同的方法可用于解决相同的问题。而每个特定程序的数学基础是已经定义的，重点放在系统的生物学知识上，这样计算机就可以作为下一个试验设计的助手而不仅仅是证明概念的方法。这门课的重要结果是使学生掌握一套计算机工具，并给他们以自信：计算机并不是无法驾驭(管理)的魔鬼。在研究项目需要时，他们拥有接触和利用新程序或系统的技能。

致谢

感谢 Wayne 州大学分子医学和遗传学中心的 D. Womble 先生，在本章中允许我们引用他的问题，并不断帮助我们成功地完成本课程，感谢我的妻子的不断支持与鼓励，本章也献给我父亲。

(张　锐　译)

参 考 文 献

[1] Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25,** 3389-3402.

[2] Pearson, W. R. (1990) Rapid and sensitive sequence comparison with FASTP and FASTA. *Meth. Enzymol.* **183,** 63-98.

[3] Quandt, K., Frech, K., Karas, H., Wingender, E., and Wermer, T. (1995) MatInd and MatInspector—new fast and versatile tools for detection of consensus matches in nucleotide sequence data. *Nucleic Acids Res.* **23,** 4878-4884.

[4] *Wisconsin Package Version 9.* 0. Genetics Computer Group (GCG), Madison, WI.

[5] Staden, R. (1996) The Staden Sequence Analysis Package. *Mol. Biotechnol.* **5,** 233-241.

[6] Benson D. A., Boguski, M. S., Lipman, D. J., Ostell. J., and Ouellette, B. F. F. (1998) GenBank. *Nucleic Acids Res.* **26,** 1-7.

26 虚拟图书馆 I：MEDLINE 搜索

Keir Reavie

26.1 引言

在过去几年中，掌握最新的分子生物学文献已变得相当容易了，因为信息资源生产者通过万维网(WWW)提供了更多信息。应用一个上网点及单个软件(一个网页浏览器)，即能使研究者及临床医师容易掌握最新的出版文献及分子生物学和分子遗传学研究动向。此外目前的进展已经允许研究者通过快速链接数据资源检索全文文章和相关的大量 DNA 和蛋白质序列的信息。

下面三章讨论可用的资源：如何利用它们按照需要来获取信息，以及提供定期更新的最新研究信息。本章重点在于 MEDLINE 数据库和如何应用国家医学图书馆(NLM)PubMed 系统来检索。第 27 章为科学引文索引(Science Citation Index，SCI)，是获取基础科学文献的重要资源；Current Contents 及其变更服务能及时地在出版时获得文献。采用文献管理软件增强和简化研究过程。第 28 章重点在获取电子期刊和主要因特网资源以使得到基金资助的机会保持更新。综观所有这些章节，主要讨论了在电子世界中保持信息的更新，尤其是在(设备)不完全情况下的上下文中，讨论了获得电子信息资源的困难，无论在技术或其他方面都是如此，以及我们对未来的期望。

26.2 MEDLINE

为搜寻生物医学文献，最易进入的和最为有用的来源是由 NLM 生产的在国家健康协会注册(NLM)的 MEDLINE。MEDLINE 是包括护士学和牙医科学文献的额外引文索引 Medicus 的 NLM 总译本。它包括文献目录编著的引文以及从美国和其他 70 多个国家出版的约 3900 种生物医学期刊来的摘要。评定标准包括 NLM[1]中可用的数据库中的期刊标题。此数据库中大约有 9 百万条记录，回顾 1996 年，当时 MEDLINE 第一次开始从打印索引 Medicus 变为电子版本格式，MEDLINE 在一周之内更新，数据库中的多数记录来自英语出版物或英文摘要。

1996 年 8 月，NLM 在 MEDLINE——PREMEDLINE 中引入一期增刊，MEDLINE 在所有记录进入 MEDLINE 之前提供存取基本的引文信息及摘要。这使出版物在

几天之内存入期刊文献。记录每天都进入 PREMEDLINE，并且每一记录设置一个 MEDLINE 惟一的识别标识符(UI)。同时所有这些记录将被全部编程入包括完全的 MEDLINE 引文因子并进入数据库。NLM 优先识别附加 PREMEDLINE 的期刊标题，并且额外的 MEDLINE 编程比其他更快。优先索引是基于对生物医学研究者们和临床医师来说相对重要的期刊而言。

NLM 将 MEDLINE 数据出售到各种不同的第三者，然后依次通过各种不同的搜索界面和各种不同的格式使数据有用，如 CD-ROMS。MEDLINE 数据库通过一个研究所图书馆，在 standalone 工作站或通过局域网使之有用，并通过第三者卖主正常买回。最多的常常是 Ovid Technologies 或 SilverPlatter。卖主也提供搜寻软件从而存取 MEDLINE 数据。1997 年 NLM 通过万维网的两种不同的界面：因特网 Grateful Med 和 PubMED 而使 MEDLINE 和 PREMEDLINE 能够自由使用。PubMED 是我们这里将关注的系统，因为它在 NLM 的国家生物技术信息中心(NCBI)得到发展，并与 NCBI 的 Entrez 数据库服务相链接。

Entrez 是由 NCBI 设计的结合存取 DNA 和蛋白序列数据库的有关生物分类学、基因组、蛋白结构信息。Entrez 也包括直接存取 MEDLINE 描述序列的文献。以此发展的自然级数包括全部 MEDLINE 数据库以及 PREMEDLINE 和 MEDLINE 记录提供与 Entrez 中的序列信息直接链接。NLM 也与出版商谈判以直接从 MEDLINE 链接到全部期刊文章。PubMed 产生的更多信息和所有其特征将注册到 NCBI 网站上[2]。

26.2.1 用 PubMed 搜寻 MEDLINE

以下有关应用 PubMed 搜寻 MEDLINE 的信息意味着基本操作指南，有关应用 PubMed 系统的详细信息通过点击 Help 链接到 PubMed 搜寻页上就可以利用了。

PubMed MEDLINE 搜寻系统(http：//www.ncbi.nlm.nih.gov/PubMed)提供两个搜寻选择：一个简单搜寻，可以链接到 PubMed 主页上；一个高级搜寻选择，适当地应用高级搜寻选择，因为它能够使你充分利用更高级的 PubMed MEDLINE 特征并改善搜寻检索。我们的主要目的是讨论 PubMed 中的更高级搜寻选择。

PubMed 主页在 Column 中提供大量的链接到左侧屏幕。Help 链接提供有关搜寻系统的详细信息。另外两个我们最关心的链接是高级搜寻和 MeSH 浏览器。由此网点直接进入 MeSH 浏览器是很有效的。我们将用其构建我们的搜寻策略。然后 PubMed 能够执行这个策略而从 MEDLINE 数据库检索到信息。当用 MeSH 浏览器搜寻 PubMed 时，结果显示更加高级的能够校正搜寻策略的搜寻选择。本章下面部分我们将讨论一些高级领域特殊的搜寻选择。

26.2.1.1 MeSH 浏览器(MeSH browser)

点击 MeSH Browser 链接进入浏览器。浏览器的目的是帮助搜寻者应用 Medical Subject Headings(MeSH)来搜寻 MEDLINE。MeSH 是应用检索文献进入数据库中的一个对照词汇。NLM 课题专家通过识别这些文章的标题来读取收入的包含在 MEDLINE 中的文献。然后进入 MeSH 浏览器来选择最适宜的专业术语来确定这些标题。这些术语先用它的索引专业术语然后到达文章或 MeSH 浏览器，并连同文章引文进入 MEDLINE。设法应用这个相似专业术语从 MEDLINE 中搜寻和进行更为精确的综合摘要信息是很重要的。其他情况下应用 MESH 能使我们避免应用专业术语同义词或缩略词的忧虑，因为所有的同义词和缩略词均注册了一个可选择的 MeSH 术语。MeSH 的应用也允许我们扩大或缩小搜寻范围，从而更容易地搜寻我们检索的相关或更专业的课题文章。

为了说明 PubMed MeSH 浏览器的应用，我们用一个例子查询:

应用因特网上的最近注册的资源文章进行序列分析，这个例子所包括的搜寻策略有两个主要概念：因特网和序列分析。在进行搜寻之前分析查询的主要概念是很重要的。然后搜寻策略在一定程度上体现所有概念而得以发展，从而检索相关信息并回答问题。在本例中搜寻将包括因特网和序列分析共同讨论的标题概念。

进入 MeSH 浏览器中的因特网术语，并点击浏览按钮，PubMed 将识别适当的 MeSH 专业术语并搜寻这个标题。浏览器解释因特网不是 MeSH，但与 MeSH 术语计算机信息网站紧密联系，MEDLINIE 中所有讨论因特网的文章将能用该术语检索，所以我们应用此 MeSH 术语检索有关因特网上的文章，MeSH 浏览器也给我们提供术语定义和一个 MeSH Tree Location(目录位置)，目录位置识别 MeSH 专业术语的序位。我们能看见计算机信息网站列于计算机系统下，局域网站作为更特殊的计算机信息网站类型列出。如果我们点击 Display(显示)链接，下一步到 MeSH 术语我们就可以检索更多的信息。

在 Detailed Display(详细显示)屏幕的近中部是次级标题列表。所有结合选择作为合格者到 MeSH 术语和帮助在更窄范围内搜寻，次级标题是一个能用来扩大标题范围的概括术语，并有助于在 Medline 中使搜寻课题范围的概括术语更专业化，它们的应用将在这部分的后面演示说明。

Detailed Display 的次级标题有两个额外的选择：Restrict Search to Major Topic headings only(仅仅是主要标题的限制搜寻)和 Do Not Explode this term (不扩展这个术语)。第一个选择仅能使我们检索我们感兴趣的并认为是讨论要点的文章(比其他更小的标题)，这将使搜寻检索范围更窄。当文章是人工索引，并且索引过程不完全客观时，需要小心选择这个选择按钮。因为一个人想的是文章的主题而另一人想的可能不是。缩小搜寻范围来检索主题有时会从检索中删除相关文章。

当选择 Major Topic headings only(只有主题标题)按钮时会缩小搜寻范围，而选择 Explode(扩展)按钮时会扩大搜寻。当扩展了 MeSH 术语时，PubMed 会自动地用 Explode 选择。扩展能使我们搜寻可选择的 MeSH 术语和下面列出的所有更多的专业术语，因此结果中会包含更多专业术语的额外文章。NLM 规则是来索引 MEDLINE 中可应用最专业的 MeSH 术语的文章。如果一篇文章在局域网站讨论序列分析，那么就可以应用局域网站进行搜寻更广义的术语，而不是应用计算机通讯网站进行搜寻。这种文章如果我们不扩展计算机通讯网站就不能在搜寻中检索到。我们需要确定在局域网中的文章是否需要包含在搜寻检索中。如果不需要，我们就关闭带有 Do Not Explode this term(不扩展这个术语)选择的 Explode 特征。

当选择了这个术语的所有选择后，我们可以点击屏幕左边包含我们查询术语的 Add(添加)按钮，这个术语就可以添加到我们的搜寻之中了。为了在我们的搜寻中添加附加术语，从本页上端选择 Enter another MeSH term to browse(进入另一MeSH)术语来浏览，并进入下一个主题：序列分析。因为我们以前用过 Detailed Display(详细显示)选项，PubMed 会自动地链接到详细显示进行 MeSH 术语序列分析，在对重要 MeSH 术语缩小或扩大进行正确选择后，我们可以点击 Add 按钮将这个术语包含在我们的搜寻之中。

当增加附加术语到 PubMed 查寻时，系统会自动地假定术语之间 AND 操作的结果，有 3 个选择：AND、OR 和 BUTNOT，这可以通过点击小窗口及下一个 Add 按钮来观看与选择，这些是 Boolean 操作时以以色列数学家 Boole 来命名的。AND 是检索两个主题同时索引的文章，OR 将检索两者之一或其他两个主题同时索引的文章；BUTNOT 除去所有的从检索地址索引的专业术语文章，在这个例子中我们将用 AND 操作。假定我们选择了两个主题的 Major MeSH terms(重要 MeSH 术语)，并打开 Explode 选择，搜寻将选择如下：

```
Computer Communication Networks [MAJR] AND Sequence Analysis[MAJR]
```

点击 Return to PubMed 按钮，到下一步搜寻策略，现在 PubMed 执行 MEDLINE 数据库搜寻。

在讨论 PubMed 显示屏幕前，让我们先看另外一个例子：

搜寻包含 X 染色体图谱的文章。

在此搜索查寻中有两个主要概念：图谱和 X 染色体。我们可以再次从 MeSH 浏览器开始。然而如果我们从图谱进入浏览器，那么它所执行的结果与上例就不同了。PubMed 不能识别图谱概念这个特殊的 MeSH 术语，并且随后屏幕上就会显示：No exact match for your term was found(你所查找的术语没有精确匹配)，从列出的术语表中我们就能够对目标进行最恰当地选择。在本例中，我们更乐于选择图谱、基因，然后点击 Browse This Term(浏览这个术语)按钮。然后 PubMed 就

引导我们查到正确的 MeSH 术语——染色体图谱，在这一点上，PubMed 查寻可以作为前一个示例。

26.2.1.2 MeSH 副标题

我们已经讨论了副标题作为一种限定 MeSH 的缩小搜寻检索的方式，点击 PubMed 的 Detailed Help(详细资料帮助)链接，就能够看到副标题的全部列表。期望 PubMed 的未来版本能够提供从副标题到专业标题定义的直接链接。在分子生物学领域和遗传学所包括的异常情况，先天的和遗传学疾病的相关副标题搜寻。例如，如果我们对遗传疾病方面，如苯丙酮酸尿的文章有兴趣讨论，那么我们首先进入苯丙酮酸尿 MeSH 浏览器，我们将会看到苯丙酮酸尿是一个 MeSH 术语，并且从详细资料显示屏幕就可以给我们提供专业性的副标题限定术语，因为苯丙酮酸尿是一种疾病，那么副标题显示就会限定相关的疾病术语，例如诊断学、病理学，或者治疗。我们可以选择我们搜寻的副标题所有组合。在本例中我们将简单地选择遗传学并要求 PubMed 将这个术语添加到我们查询之中，假定我们选择的 PubMed 检索的这些文章是 MeSH 的重要主题，那么搜寻检索就会显示如下:

```
Phenylketonuria/genetics[MAJR]
```

26.2.1.3 修改与限制搜寻结果

一旦 PubMed 执行了 MEDLINE 数据库的搜索查询，那么将显示 PubMed 查询屏幕。最顶部是搜索查询，接着一个按钮就会显示大量检索的文章，我们可以点击这个按钮，就开始显示默认同时 20 个索引文献。PubMed 会自动地显示大量追溯到 1966 年的可利用的文献。我们可以改变 Entrez Date Limit(日期限定)选择，来限制显示的文献数量。选择范围是从 30 天到无限(追溯到 1966 年)。

查询结果屏幕的下一部分标题是 Add Term(s) to Query(将术语添加到查询之中)。这部分能够将额外的术语添加到我们的搜寻之中，从而使该术语更加专业化，通过点击查寻区域窗口内的光标，你可以看到我们能够在 MEDLINE 数据库内查到专业域，它包括 MeSH 和主要的 MeSH 区域。不幸的是，在此阶段我们不能返回到 MeSH Browser 添加另外的 MeSH 术语，但我们可以把 MeSH 术语放到 Add Term(s) to Query，并在搜寻窗中选择 MeSH 浏览器。这里假定查询者知道他们希望搜寻的 MeSH 术语。如果我点击搜寻，PubMed 将自动地把这个 AND 术语并入到当前查寻条，我们也可能从 Search Mode(搜寻格式)窗口选择目录框 List Terms 按钮。这使 PubMed 首先为查寻者在选择运行此搜索前可以提供适当的术语。查寻者在不清楚他们想找的 MeSH 术语确切用法的情况下，这是一有用的特征。

当浏览我们的检索并限定检索结果之前，我们或许想做一些标准化工作。其中一些限定因素包括缩小英语文献的检索范围，如研究人类和综述性的文章。

为了保证这项工作的正确性，我们需要在正确的 MEDLINE 数据库范围内搜寻这些术语。英语应该在 Language(语言)域搜寻，而综述应该在 Publication Type(出版类型)域中搜寻。可利用的出版类型的完全表格由点击 PubMed DeTailed Help 链接来列表。进入这个术语然后从搜寻窗口选择恰当的范围。人类研究的搜寻要稍微复杂些，Human 是一个 MeSH 术语，但在检索中主要作为一个检验标签，检验标签[3]的其他类型常用来标识地理区域或年龄群体的研究。因为标签比较稀少，则从来都不作为主要的 MeSH 术语进行搜寻，即使作为主要的索引术语，它们也可能是例外。为了缩小人类研究的检索，则在 MeSH 域中搜寻 Human。希望 PubMed 搜寻系统的未来版本包括更多的搜寻选择，并且将用这里讨论的标准搜寻限定提供更便捷的方式。

仔细地应用限定英语是非常重要的，如果你的搜寻需要综合，请记住许多重要文献是除了英语以外的其他语言出版物，当仅仅缩小英语文献范围时，许多重要的文献将会丢失。但是以其他语言出版的文献在 MEDLINE 中均有英文摘要。

26.2.1.4 在 PubMed 中显示检索结果

从 PubMed 搜寻检索屏幕上显示搜寻的结果，选择一个适当的 Entrez Date Limit 并点击搜寻中的大量检索文献按钮。下一个屏幕将显示检索中的前 20 个引文，引文可利用的详细信息通过点击作者姓名链接来完成。每篇引文同时提供 Related Articles(相关文献)的链接，这个链接可以用标题中、摘要中和 MeSH 域中的相似术语索引到紧密相关的文献。这个特点对于一旦发现一篇文章与我们所查寻的文章密切相关时是十分有用的。然后相似的文章也能容易地索引到。

标准显示选择是提供有用的基础文献目录信息和一个摘要的 Abstract Report(摘要报告)，我们可以通过点击每一个左边的同时显示的引文群体 ，一旦从 Display 窗口点击 Display 按钮，就进行了选择。此外，Abstract Report 和 MEDLINE Report 格式的引文是为了将下载的结果输入到各种文献目录管理软包装(参见第 27 章)。PubMed 也从相关信息显示链接的其他 Entrez 数据库提供选择。每篇引文要进行全屏显示方式，那么在上部有一个额外的显示链接按钮，这样能够比较快捷地直接链接到 Entrez 数据库文献的完全文本。这些链接按钮包括：蛋白质——SwissProt、PIR、PRF、PDB 和从 DNA 数据库中翻译出的蛋白质序列；DNA——来自 GenBank、EMBL 和 DDBJ 中的 DNA 序列；OMIM ——来自人的孟德尔遗传规律的在线信息；并且链接到出版者站点或完全的文本。依靠 NLM 赞同的每个出版者在单项的基础上进行完全文本存取是有效(用)的，PubMed 用户授权存取完全的文本。完全文本存取的多种情况在第 28 章予以讨论。

用你的 Web 浏览器的打印选择或从文本中下载或用 PubMed 全屏显示底部的 Save 选择进行 HTML 格式化就可以很方便地打印出引文。如果你有意将你的

搜寻结果输入文献目录管理软件，那么引文就将以 MEDLINE 格式显示并以文本文件下载。

26.2.1.5 保存搜寻结果

PubMed 搜寻很容易执行并且能够用 Simple Search 屏幕保存，以备将来使用。这项工作，搜寻者必须在没有 MeSH 浏览器帮助的情况下手工进入搜寻。

如果想要保存和返回这部分我们首次创建的搜寻，我们只要进入 Simple Search 屏幕，并点击 Search 按钮就可以进入：`Computer Communication Networks[MAJR]AND Sequence Analysis [MAJR]`。

用你的 Web 浏览器标记书签搜寻检索页就能够将你的搜寻保存下来以备将来之用。在任何时候想重新查寻，只要简单地点击标记书签便可。然而不幸的是，当你重新执行检索时，又要搜寻整个 MEDLINE 数据库。如果你仅仅想要检索当前信息，进入你最新搜寻的 MEDLINE，则最好在你的搜寻策略中进行时间限定。

不用 MeSH 浏览器而要提高你的搜寻策略就要求具有 PubMed 搜寻命令语言的知识。用这种语言得到的信息能够从任何 PubMed 屏幕上点击 Help 找到手动存取的在线信息。

26.2.1.6 Loansome Doc

Loansome Doc 是 NLM 提供的能使 PubMed 搜寻者在他们选择的图书馆进行跟踪搜寻结果的一种服务器，因此图书馆就能向使用者依次检索和传递文献全文。应该注意到地方图书馆提供这种服务，并且依靠这种图书馆可以进行文件传输管理。用 Loansome Doc 进行更多的信息搜寻要定位于 NLM 的网址[4]。在美国列出的图书馆将能接受和填写 Loansome Doc 从而能够在国家医学图书馆网络(NN/LM)的网站(`http://www.nnlm.nlm.nih.gov/`)找到。

在 PubMed 显示屏幕上，Order 按钮直接定位于 Display 按钮下方。文献能够通过点击 box 左边的引文然后点击 Order 进行排列而选择。Loansome Doc 要求用户用 ID 和 password(口令)进行登录。如果你是 Loansome Doc 的首次用户，则要求你注册这种服务并分派给你一个 ID 和口令以备将来之用。你还需要一个图书馆 ID 来完成注册，图书馆 ID 告诉 Loansome 排列要求的路径，图书馆 ID 可以从 NN/LM 网站选用或者通过与你的地方图书馆联系而得到。

附录

以下列出的是本章讨论的万维网的主要网址：

National Institutes of Health: `http://www.nih.gov/`

National Library of Medicine: `http://www.nlm.nih.gov/`

PubMed MEDLINE: http://www.ncbi.nlm.nih.gov/PubMed/

(张　锐　译)

参　考　文　献

[1] National Library of Medicine (August 27, 1998), *Journal Selection for Index Medicus/MEDLINE,* National Library of Medicine [3 pages], http://www.nlm.nih.gov/pubs/factsheets/jsel.html.

[2] National Center for Biotechnology Information (January 9, 1998), *The NLM PubMed Project,* National Library of Medicine [3 pages], http://www.ncbi.nlm.nih.gov/PubMed/overview.html.

[3] National Library of Medicine, Medical Subject Headings Section (1998), *Medical Subject Headings, Annotated Alphabetic List,* 1999, National Library of Medicine, Bethesda, MD.

[4] National Library of Medicine (June 2, 1998), *Loansome Doc,* National Library of Medicine [2 pages], http://www.nlm.nih.gov/pubs/factsheets/loansome_doc.html.

27 虚拟图书馆II：科学引文索引和更新通告服务

Keir Reavie

27.1 引言

接上一章，本章继续讨论如何在因特网检索生物医学文献。内容主要有：如何在 MEDLINE 链接科学引文索引(Science Citation Index, SCI)，如何利用更新通告服务(current awareness service)，如 Current Contents 数据库，以及目录通知服务(tables-of-contents alerting service)获取最新的已出版的文献。本章最后还讨论了参考文献管理软件以及如何把它结合到查询过程中。

27.2 科学引文索引

科学引文索引(SCI)是多学科的书目数据库，由科学信息所(Institute for Scientific Information，ISI)出版发行。SCI 提供了一种检索 MEDLINE 数据库文献的重要途径，特别是有关基础科学研究的出版物，而且它还包括许多 MEDLINE 没有收录的生物医学期刊。SCI online 内容与印刷版 SCI 一致，数据库中每一条记录都包括文章引用的文献，研究人员可以通过 SCI 检索某一被引用文献的作者或文章。通过检索被引文献，研究者可以了解有关某一主题更多更新的资料，而这些资料有时通过主题检索是无法检索到的。SCI 最大的优点是，研究者能够回溯检索某一研究活动历史动态，而且能分析某一已发表研究(论文)的影响力。

SCI 覆盖科学技术所有学科，大约收录 3500 种科学技术方面的权威期刊。网络版 SCI(ISI's Web of Science)，收录 5600 种期刊，可以回溯检索到 1974 年。电子版 SCI 可通过多种途径检索，例如，光盘版，每月更新；磁带版，每周更新。每周更新的数据包括 17 000 最新文章以及大约 300 000 最新引文，其中 70%的文献带英文摘要。有关 SCI 的详细说明、覆盖的学科及使用方法参见 ISI's Web 主页面[1]。

27.2.1 通过 Web of Science 检索 SCI

下面将讨论在 Web of Science 上查询 SCI 的一些技巧。在网上检索 SCI 不像 PubMed MEDLINE 那么复杂，这主要是因为 SCI 没有像 MEDLINE 的 MeSH 那样

的受控词表，也没有 Entrez 数据库中便捷的分子和遗传方面的信息。有关 Web of Science 详细资料可通过点击 Web of Science 主页 Help 按钮获取。Web of Science 检索 SCI 不像 PubMed MEDLINE 是免费的，只有与 ISI 签署使用协议的会员才允许使用。

Web of Science 主页 http://webofscience.com/提供两种检索 SCI 页面：Quick Search 和 Full Search。这里只介绍 Full Search。首先点击 Full Search，进入数据库选择屏幕，需要选择所需的数据库。选择检索数据库的权限取决于购买的数据库。可选择的数据库有：Entire database、This week's update、Latest 2 weeks、Latest 4 weeks，也可以选择检索某一年的 SCI 数据。同时 SCI 还提供两种方法进行检索：General Search 和 Cited Ref Search。General Search 提供主题检索途径，Cited Ref Search 通过被引文献的作者或出版物进行检索。

27.2.1.1 主题检索

选择 General Search 功能选项进入 Web of Science 检索界面。在屏幕中间位置有 4 种检索选项：Topics、Authors、Source Title(检索特定杂志)和 Address(从机构入手检索出版物)。Address 适合从某一机构入手检索出版物，而且通过分析该出版物的被引率来评价该机构的科研水平。

前面已经提到，SCI 没有受控词表，所以主题检索需要认真仔细，以免漏检。但 SCI 有关键词检索功能，这些关键词有些是文章作者提供的，有些是 ISI 增添的关键词(keywords plus)。这些增添的关键词在被引文献中出现频率很高，但在文章标题中没有出现的词或短语。主题检索时将自动对这些关键词、文献题目及文摘进行检索。由于 SCI 没有受控语言，检索者在制定检索策略时应考虑使用同义词和缩写词，因为不同作者在描述同一主题时可能用不同的关键词。同样，截词检索也很重要，“*”是截词符号，用来检索词根相同但词尾不同的词。例如，Genetic*能检索到 Genetic、Genetics、Genetically 等。

Web of Science 检索界面不很灵活，意味着一旦检索已经执行，不容易修改。在设计检索策略时，检索者必须制定一个全面、精确的检索式。例如，我们检索第 26 章 MEDLINE 中曾检索过的“ the genetic aspects of phenylketonuria”。在 Web of Science 检索式应该如下制定：

```
phenylketonuria AND (genetic* OR molecular biology)
```

检索这一主题，这个检索式不是最佳选择，但它是检索 Web of Science 的一个实证。Web of Science 可以使用布尔逻辑运算符 AND、OR 和 NOT。用括号将 genetic * OR molecular biology 括起来，检索过程是 phenylketonuria 和 genetic，或者是 phenylketonuria 和 molecular biology，而不是先检索 phenylketonuria 和 genetic，然后再与 molecular biology 进行逻辑“或”的检索。因为在布尔逻辑运算中，AND 运算优先于 OR 运算，因此括号的使用非常重要。

在 General Search 界面底部是 limit 和 sort 两个限定检索的选项，这里能进行限定的只是语种，将语种限定在英语。记住，如果想进行全面检索，则不需要进行语种的限定，这也许会将非英文发表的非常重要的出版物漏检。如果没有指定，Web of Science 将自动按最新数据进行检索。按 Search 功能按钮，将执行检索。

27.2.1.2 检索结果显示

一旦检索完成，Web of Science 立即显示出检索结果的前 10 条记录。检索结果显示屏幕与 PubMed 结果显示屏幕相似。点击文章题目将显示详细的检索结果，包括文摘、被引文献。点击记录左侧显示、打印及存盘按钮，可以进行全文显示、打印及存盘。如果要对显示结果进行选择标记，则在选择的记录前标记，然后点击 Submit 功能按钮，将选择的记录添加到 Marked List。如果你没有按 Submit，选择的记录将被删除。如果所有相关文献都已选择完毕，按 Marked List 功能按钮，将看到所有选择的记录，然后进行打印或存盘。

Marked List 屏提供选择字段选项，以及进行打印或存盘的特定排序方式选择界面。有 3 个用于输出选项的功能按钮：Format for Printing，是在 Web 浏览器中选择打印记录的格式；Save to File，是记录的存盘；Export，是将检索结果自动输出到一个书目管理软件包(见 27.5 节)。

27.2.1.3 引文检索

SCI 引文检索是通过文章的被引作者或被引出版物来检索某一主题的更多最近文献。点击 Web of Science 主页顶部的 Cited Ref Search 功能按钮则进入引文检索界面。

引文检索可检索 Cited Author(被引作者)，还可在检索界面的 Cited Work 对话框区输入杂志缩写，或在 Cited Year 对话框输入出版物的年来检索被引作者的著作。注意：该系统要求输入杂志的缩写。点击位于 Cited Work 上的 List 按钮则可进入杂志缩写目录。作者姓名的书写要求是：先是姓，空一格，然后是第一个及第二个名的首字母的大写。如果第二个名不知，则可在第一个名后加截词符“*”。因为一些作者的第一个和第二个名的首字母相同，这样用截词检索的结果将会包括一些错误结果。如检索 Stephen Krawetz，在 Cited Author 框内输入 `krawetz s*`。如果已知作者的中间名字是 A，则输入 `krawetz sa`。点击 Lookup 按钮，则开始检索该作者的被引情况。

引文检索结果显示：在上述检索中，被引作者的检索结果每屏显示 10 条记录，记录左侧是被引频率，位于 Hits 之下。如要检索引用某一出版物的全部文献，在屏幕左侧相应对话框中输入检索词，然后点击 Search 按钮。在显示屏幕也有像在 General Search 中的限定和排序按钮，在执行检索前需进行修正。

27.2.1.4 检索保存

Web of Science 允许检索用户保存检索式，以便以后在数据更新后进行检索。这一功能在 General Search 和 Cited Ref Search 都支持。首先进入检索界面，点击 Save Query 按钮。Web of Science 立即将检索式以指定的文件名存入指定磁盘，文件的扩展名为“.cgi”。要执行一个已保存的检索式，首先要将它读取出来。这通过选择 Web 浏览器中的 File 菜单中的 Open File 命令。要打开这一文件，必须先进入 Web 浏览器，然后才能进行检索。检索方法与其他检索相同。

27.3 Current Contents

Current Contents(CC)是由 ISI 发行的全球第一个现刊题录及文献快讯数据库。CC 的内容来自最新出版的有关科学、社会科学、技术、艺术及人文科学的 7000 多种杂志和 2000 多种书。印刷版 CC 只有目录，电子版 CC 包括每篇文章的详细书目信息及英文摘要，同时还提供有关作者的信息。

CC 每周更新，共分以下 7 个专题出版：

(1) 农业、生物和环境科学

(2) 艺术人文

(3) 临床医学

(4) 工程、计算机和技术

(5) 生命科学

(6) 物理学、化学和地球科学

(7) 社会和行为科学

每个专题之间可能会有重复。如要广泛检索分子生物学的文献可以使用(1)、(3)、(4)、(5)和(6)专题。因为 CC 报道最新信息，所以不能回溯检索很长时间以前的内容，只能检索到最新内容。回溯检索需要在 MEDLINE 和 SCI 上运行。

数据库以多种格式存在：ftp 传输、软盘、CD-ROM，都可从 ISI 或者中间商购买。ISI 现在提供网上检索服务(Current Contents Connect，CCC)，详细说明及使用方式见 ISI 网站[2]。

27.3.1 检索 Current Contents Connect(CCC)

这里只简单介绍 CCC 的使用方法，如要了解详细情况，请点击 Current Contents Connect 主页的 Help 按钮。Current Contents Connect 是收费数据库，只有授权用户才能使用。

如果链接到 CCC(`http://www.isicc.com/`)，点击 Start 功能钮，进入 Search Limits 屏幕，在检索前需要在此选择搜索限定。屏幕左侧是 Current Content

Editions 编辑框，均处于开启状态。如不需要某一专题，则点击该专题编辑框中左侧的复选框关闭之。如检索分子生物学的文献，则不需要社会和行为科学、艺术及人文科学家专题。屏幕右侧是 File Depth 选项，它提示你选择了哪一数据库：Latest week、Latest 4 weeks、Latest 6 months、Extended file(最近 2 年文档)。一旦选择好后，点击 Submit Limit Changes 按钮，进入检索状态。

CCC 检索方法与 Web of Science 上的 SCI 相似，CCC 屏幕右上侧的信息告诉用户所选择的专题及数据库，屏幕中间及左侧是 Topic/Subject(TS)窗口，点击下拉菜单，则出现 CCC 可检字段选项。正常状态下，选择好字段，然后输入一个检索式。如检索与前边介绍的 SCI 相同的检索式，先选择 Topic/Subject，然后输入：

```
phenylketonuria and (genetic* or molecular biology)
```

点击 Search 按钮执行检索。与 SCI 一样，CC 没有受控词表，只有关键词，这些关键词有些是作者提供的，有些是数据库根据内容增添的(见 27.2.1.1 节)。

与 Web of Science 马上给出检索结果不同，CCC 执行完一个检索后，重新回到检索界面。检索式出现在屏幕底部，并给出命中文献数。点击眼状图标，则出现前 10 条命中文献。在该界面可以进行新的检索，检索式与第一个检索式罗列在一起。该屏只能容纳 10 个检索式，如果超过 10 个，最早的检索式将被删除。检索式可以进行组配检索，例如，在可检字段的语种字段内选择 `English`，生成一个检索式 2，检索式 2 可以与前面的检索式 1 组配检索。点击 Search Field 菜单中 Combine Sets，在检索框中输入 `1 AND 2` 即可。记住，在进行组配检索前要选择 Combine Sets，否则 CCC 将对所有含数字 1 和 2 的记录进行检索。

27.3.1.1 检索结果显示

CCC 检索结果显示与 Web of Science 基本相同，不同的是 CCC 只有 Mark All 一个选项，选择后对所有检索结果进行保存、打印。如果只想选其中某一文献，需点击该文献题目，进入全文浏览状态，在全文浏览状态下点击位于左上部的按钮。CCC 已意识到这一不足，正在改进。

如果已标记好选择记录，点击 List 按钮则会看到全部所选记录。List 界面也提供了 Sort、Print、Download、Export 选项。其中 Export Format 选项包括 Request-a-Print File 格式。Request-a-Print File 使输出格式可以打印到复制索取卡，该卡可向 ISI 购买，并邮寄给作者要求复制。还可以输出到 Procite 和 Reference Manager，这是两个书目管理软件包，点击 Export to Procite/Reference Manager 按钮即可(见 27.5 节)。

ISI 还提供直接从其服务器订阅文献的服务。注意，版权费加上邮寄费，订阅费用会很贵。

27.3.1.2 检索式存储

由于 CC 报道的是最新书目信息，因此将检索式保存起来，以便以后每周更新检索结果。CCC 保存检索式的方法同 Web of Science。在 Search 界面点击 Save Session 按钮，CCC 开辟了一个叫 Current Content Connect Profile 窗口，完成的检索将列在此窗口。点击 Save Profile 按钮，检索式将存入指定磁盘(见 27.2.1.4 节)。以后如需打开该检索式，点击 Web 浏览器 File 中 Run Profile 即可。CCC 在运行保存的检索式前，提示用户是否改变以前设置的限定条件。这是在 Search Limits 屏幕进行的，用户可以重新选择新的检索专题，也可以选择 File Depth，如 Latest Week，来检索最近一周的更新数据。

27.4 及时通知服务

该服务器最近为读者提供最新的书目快讯。它最大的优点是当有新的更新数据时及时通知用户，为用户减去了定期记住并登录进行检索的负担。

27.4.1 ISI Table of Contents Corporate Alerting Service (ISI 目录联合公告服务)

ISI Table of Contents Corporate Alerting Service(ISI TOC)是通过电子邮件提供服务的。用户从 ISI 数据库大约 8000 多种杂志中选择、订阅，通过电子邮件获取文摘。详细情况参见 ISI 的 Alerting Services Site[3]。

ISI TOC 使用起来非常容易。先给 ISI 的一个特殊的 e-mail 账户发送要求登记信息，半小时内就会给申请者发来 200 多个主题领域表单。申请者在感兴趣的主题左侧标记“X”，然后把表返回给 ISI。半小时之后，ISI 向用户发所选主题的杂志目录。用户再选择所需杂志，再返回至 ISI。所有选择的杂志名称记录在申请者的文件内。当选择的杂志新出版后，ISI TOC 用户将从邮箱获得杂志目录内容。

ISI TOC 另一个特点是，用户可以在目录上选择所需文献，并且提交给当地图书馆或返回到 ISI，索取原文。

27.5 文献管理软件

文献管理软件(BMS)在 27.2 节中几次提及。BMS 有两个主要功能：①使用户能对某一专题创建并维护个人文献数据库；②为文章脚注和尾注以及参考文献提供一种格式，便于出版。

BMS 常用的软件包有 Reference Manager(RM)、Procite 及 Endnote，它们的特

点及费用基本相同。

RM 最初是为生物医学研究者设计的，根据文章所呈送出版物来形成某些杂志的参考文献信息格式。

RM 和 Procite 是由 Research Information Systems(RIS；http://www.risinc.com/)开发的，Endnote 是由 Niles and Associates(http://www.niles.com/)开发的。RIS 和 Niles and Associates 现都隶属于 ISI。本书编写过程中，它们正在组成一个新的公司，ResearchSoft。RIS 已经承诺，虽然长远来看同一家公司不可能出售非常相似的 3 种产品，但 RIS 还会延续这 3 种软件包的技术支持和产品的开发。

RM、Procite 和 Endnote 都有从 MEDLINE、SCI 及 CC 自动下载文献的功能。事实上，Web of Science 的 SCI 和 Current Contents Connect 的 CC 检索界面上都有选项，使用户直接从万维网下载数据输入 RM 和 Procite。这需要从 RIS 下载一个软件，很快 Endnote 也会拥有该项功能。

Stigleman[4]讨论和比较了这 3 个 BMS 软件包的主要特点，这篇综述时间上稍微有些早，但它综述了软件的功能及重要信息，这对于要购买软件包的用户非常有价值。

附录

本章涉及的网站：

Institutes for Scientific Information http://www.isinet.com/

Science Citation Index on the Web of Science http://webofscience.com/

Current Contents Connect http://www.isicc.com/

Research Information Systems http://www.risinc.com/

Niles and Associates http://www.niles.com/

(逄大欣 译)

参 考 文 献

[1] Institute for Scientific Information (March 30, 1999) *Science Citation Index Database,* Institute for Scientific Information [6 pages] http://www.isinet.com/prodserv/citation/citsci.html.

[2] Institute for Scientific Information (January 27, 1999) *Current Contents General Information,* Institute for Scientific Information [5 pages] http://www.isinet.com/prodserv/cc/cchp.html.

[3] Institute for Scientific Information (March 30, 1999) *ISI Alerting Services,* Institute for Scientific Information [3 pages] http://www.isinet.com/prodserv/ias/ca.html.

[4] Stigleman, S. (1996) Bibliography programs do Windows, *Database* **19**, 57-66.

28 虚拟图书馆 III：电子期刊、赠款、基金资助信息

Keir Reavie

28.1 引言

本章内容主要讨论如何在因特网获取电子期刊，以及遇到的问题。本章最后部分介绍了如何利用因特网获得资金资助。

28.2 电子期刊

在因特网浏览杂志的全文是一个比较新的事物。如何方便、容易地得到全文，许多细节问题，如权限、价格、技术问题等都急需解决。杂志发行有两种渠道：一是商业出版，另一个是由机构或协会出版发行。在因特网上获取这两种渠道发行的期刊全文都有一些相似的特点和政策，但在阅读权限、价格等方面都有很多不同。一般一种杂志联上网，它能提供目录、文摘或全文。如果有全文，全文的格式有以下两种：HTML 和 PDF 格式。很多杂志提供这两种格式。PDF 在很多方面更有优势，以下将详细讨论。

访问全文的价格是一个很重要的问题，不同杂志、出版社之间有很大差异。由机构或协会出版的电子期刊会向该协会会员免费提供全文。它也向订购该杂志印刷版的用户免费，或收取少许费用。而由商业出版社出版的杂志则向用户收取很高的费用。最近有些商业出版社也采取对订购印刷版的用户收取很少的费用，但许多商业出版社仍旧收费很高。许多迹象表明，以后的趋势是在印刷版费用之上另收联机阅览费用。一些大的出版机构只以专辑的形式提供全文——以高于印刷版两倍的价格购买电子版杂志。这种情况一般只限于拥有印刷版的用户。但是，有时用户可以多花一些钱购买更多的杂志，或这些杂志中某些特定的文章。一些大的出版社会通过网站通知费用的变化。一般学术协会这样的大机构能够承受这笔费用，它们在以后也会为它们的教员和学生订购这些杂志。

目前获取电子版全文的渠道是相当混乱的，许多重要问题需要在相当长时间才能解决。现在获得电子期刊全文浏览权限最简单的途径就是在网上查找杂志，并研究如何获得所需的杂志。如费用是多少？如果有印刷版，如何获取电

子版全文？有时在订阅印刷版的同时，要求用户订购电子版，这与以上提及的情形正好相反。

在订阅全文电子版期刊时，除了考虑费用问题，还应注意的是：如何控制对全文的访问？该控制方式会对主要用户的访问权限有影响吗？如果以后不再订阅，还能访问以前的期刊吗？因为如果印刷版杂志订阅取消，会保留以前的杂志，这在电子版也能做到吗？有关访问的政策和技术问题涉及用户与出版社双方。在因特网获取全文主要通过两种方式：一是注册 ID(用户名)和 password(密码)，二是通过指定的因特网地址，通常会分配给一个计算机地址段范围，属于某一团体、机构或组织。登录 ID 方式需要全面管理分配给所有用户相同的 ID 和 password，或者给每一个用户分配不同的 ID 和 password。对于通过因特网地址访问电子期刊的用户来说可能会遇到一个问题，就是有些主要用户的工作位置落在指定的访问地址以外。出版商可能不愿意在没有收取额外费用的情况下扩大访问地址范围，因为他们认为会存在附加的用户群。那么对于必须拨号上网的用户来讲，如果其地址和单位机构提供的访问电子期刊的地址不同，他们的访问就会受到限制。从策略上来说，如果一个协会或机构的成员的办公室不在机构主楼，他就没有权力进入该机构主楼了吗？在许多大研究机构都存在这一问题。因为各个科室可能会分散在不同的地方，并且对于研究人员来说，能够访问电子期刊是最重要的。

目前用户必须从不同的万维网位置进入不同的电子期刊，一个有效的办法就是将用户检索界面与访问点(access point)结合在一起。这正在 4 个层面发生变化：出版商、机构或协会、检索网站(如 PubMed)及一个第三方机构。后者包括像 Stanford 大学 Highwire Press 这样的出版商，这在下一部分内容里的电子出版实例中将详细讨论。

28.2.1 电子期刊实例

为说明访问电子期刊中的一些问题，我们要举出一些在线杂志的例子。例子不是很全面，读者在订阅其他杂志时可能还会遇到其他问题。

28.2.1.1 Nature

从 1998 年 9 月开始，个人订户可免费浏览 Nature、Nature Genetics 和 Nature Medicine(`http://www.nature.com/`)、Nature Neuroscience、Nature Biotechnology、Nature Structural Biology 全文，但团体用户不能。研究人员要想浏览电子版期刊，必须申请个人账户。

28.2.1.2 Science

电子版全文 Science(`http://www.sciencemag.org/`)向个人用户收费 12 美元。只允许并发一个用户，由所使用的计算机控制。协会会员图书馆每台

计算机收费 25 美元。采用基于全职雇员(和教育机构的学生数目)和附属地址数目的机构可以按网址范围访问。(译者注：该杂志全文在国内大多数地址可以自由访问。)

28.2.1.3 Proceedings of the National Academy of Science(PNAS)

Proceedings of the National Academy of Science(PNAS)(http://www.pnas.org/)电子版要求在印刷版的基础上另附费用。PNAS 是基础医学核心期刊之一，由 Stanford 大学 Highwire Press 发行(http://highwire.stanford.edu)。许多学校和科研机构出版的期刊都与 Highwire Press 签订协议，以 HTML 和 PDF 格式向 Highwire Press 提供重要的生物医学期刊。Highwire Press 的网站网集了所有签约的期刊，如 PNAS，每种期刊的费用不同，都需用户与期刊的出版商签订。

PNAS 需要申请 ID 和 password，如果是团体用户，需向 PNAS 出版商提供该机构的因特网 IP。Highwire 上每种期刊的个人账户申请程序均相同，只不过是向不同的出版商申请。Highwire 在这里起的作用是订户与出版商的中介。

28.2.1.4 EMBO Journal

EMBO Journal 电子版也是通过 Highwire 来订阅，最近它免费向印刷版订户提供电子版全文。

28.2.1.5 Journal of Biological Chemistry(JBC)

Journal of Biological Chemistry(JBC)(http://www.jbc.org/)是最早可在网上浏览全文的期刊之一，通过 Highwire 订阅，它对美国生物化学和分子生物学学会会员是免费提供的。团体用户的订阅费用比订纸版的费用高。(译者注：该杂志全文在国内大多数地址可以自由访问。)

28.2.1.6 Cell

印刷版 Cell 用户可免费使用其电子版(http://www.cell.com/)，团体用户订阅的费用比订纸版的费用高。这笔高出的费用取决于该机构规模大小与联机最大并发用户。学术机构用户还可同时阅览 Cell 出版社的其他杂志，包括：Immunity、Neuron 和 Molecular Cell。

28.2.1.7 Molecular Medicine Today

Molecular Medicine Today (http://www.elsevier.com.locate/molmed/)是由 Elsevier 发行。该出版社的出版物可通过其 Science Direct 服务获得。由于其需要对 Science Direct 系统的初始化费用，所以一般只向团体用户提供，费用变得很高。

28.2.1.8 New England Journal of Medicine(NEJM)

New England Journal of Medicine(NEJM)(http://www.nejm.org/)免费为订阅纸版用户提供电子版全文，包括个人用户和团体用户。ID 和 password 可在 NEJM 网站上申请。团体用户用此方法会遇到问题，因为他们也要使用一个 ID 和 password，最好是通过因特网网站登录，以便机构内所有用户不必每次都得记住 ID 和 password 才能使用电子期刊。如果每种期刊都有一个 ID 和 password 记起来更是麻烦。

28.2.2 用 Adobe Acrobat Reader 浏览 PDF 文件

电子期刊主要的格式有 HTML 和 PDF。HTML 格式对于此类出版物来说有问题，因为先浏览的全文 HTML 文件主要是文本和图片、表格和图表的缩略图或者小图片。如要阅读这些图表，必须点击这些图表的链接，在另一个 Web 页面中加载完整图像。如果要将文章打印出来，就得先打印文本部分，然后再将每个图表分别打印出来。另外，HTML 格式还在打印时不能正确分页。而 PDF 格式与印刷版文章的格式完全相同。

浏览 PDF 格式文献必须用从 Adobe 公司免费下载的 Adobe Acrobat Reader(http://www.adobe.com/)，而且许多用 PDF 格式提供文献的网站都有与 Adobe 的链接。安装之后，Web 浏览器在浏览文献时将自动打开 PDF 格式的文件。而且 Adobe Acrobat 也只能打开 PDF 的文件，它没有编辑的功能。如果打印带图表的 PDF 文件，最好有一台好的激光打印机。

28.3 拨款和基金资助信息

因特网上有两个网站提供拨款和基金资助的信息：①Nation Institutes of Health (NIH) 的 Grants and Funding(http://www.nih.gov/grants/)；②National Science Foundation(NSF)的 Grants and Awards(http://www.nsf.gov/home/grants.htm)。两个网站都能进行浏览和检索，以便用户及时了解基金资助的信息及申请要求。

28.3.1 Nation Institutes of Health(国家卫生研究所)

了解 NIH 基金信息可通过两种方式：一种是通过 NIH Office of Extramural Research(OER)(http://www.nih.gov/grants/oer.html)，另一种是通过 NIH Guide。Department of Health and Human Services(包括 NIH)资助的信息都可在 CRISP(Computer Retrieval of Information on Scientific Projects)数据库中查到。NIH 下设的各个研究机构也提供资金资助的信息，可在其网站上检索到。NIH 主页上

链接了各个研究机构的网站(http://www.nih.gov/)。

28.3.1.1 检索 Office of Extramural Research(OER)

OER 资金资助网站提供检索 NIH 资助项目的服务(http://www.nih.gov/grants/search.htm)，该网站页面有一检索框，可直接输入检索词或短语。该网站没有像 MeSH 一样的主题词表，所以在进行检索时要小心使用术语、同义词及缩写。但它允许使用操作符号<thesaurus>，在任一词条前采用此操作符号会得到含有该词条同义词的结果。检索提示位于页面的底部。如果检索时遇到问题，可点击 Help 按钮。检索时可以使用布尔逻辑运算符 AND、OR、NOT 以及截词符“*”。例如，检索有关丙酮酸尿症遗传方面研究资助情况的检索式如下：

```
phenylketonuria AND genetic*
```

点击 Search 按钮执行进行检索，检索结果马上显示在屏幕上。检索结果包括题目和摘要的前两行，点击题目浏览详细内容。注意，OER 检索系统不能检索 NIH Guide to Grants and Contracts 上的 Requests for Applications and Program Announ- cements。

28.3.1.2 NIH Guide to Grants and Contracts(NIH 基金与合同指南)

NIH Guide to Grants and Contracts 上的 Requests for Applications and Program Announcements 以印刷版和电子版的形式每周发布一次(http://www.nih.gov/grants/guide/index.html)。NIH Web 页面提供每周可浏览本年发布的信息的链接，也可以回溯检索到 1992 年来资助的信息。检索框位于窗口的中部，使用的检索系统与 OER 相同，因此具有相同的特点。点击搜索行右侧的 Help 按钮可获得检索帮助。

NIH Guide Listserv：向 Listserv@list.nih.gov 发一个 e-mail 加入 NIH Guide Listserv，就可以定期每周得到 NIH 资助项目的信息，e-mail 内容是：subscribe NIHTOC-L your name，NIH Listserv 系统将你加入 NIH Guide List。几分钟后，你会收到信息，确认你是否加入，并指导你如何使用该列表及如何取消。一旦注册成功，NIH Guide 将每周给你的 e-mail 账户发送新出版的信息。有关 NIH Guide Listserv 的信息可以在线获得[1]。

28.3.1.3 CRISP

CRISP 数据库(http://www-commons.cit.nih.gov/crisp)提供 Department of Health and Human Services 资助的研究项目和计划的信息。它的主要信息来源是 NIH，但还包括 Centers for Disease Control and Prevention(疾病预防和控制中心)、Food and Drug Administration(食品和药物管理局)、Health Resources and

Services Administration(卫生资源和服务部)、Agency for Health Care Policy and Research(健康照护政策与研究机构)。从 CRISP 可以检索到谁得到了某研究项目的资助，以及是谁资助了这些资金。网络版 CRISP 是最近推出的，存在许多缺点和不足，影响到快速和有效的搜索。下面将会提到一些方面。

检索 CRISP：CRISP 有两个数据库：Current Award Information 和 Historical Award Information。如果想得到更多最新的资助信息，选择 Current Award Information。CRISP 提供两种检索方式：Basic Query Form(基本检索)和 Advanced Query Form(高级检索)。如果有什么问题可点击位于屏幕右上部的 Help 图标。Advanced Query Form 允许使用布尔逻辑运算符。例如，还是检索丙酮酸尿症遗传方面研究资助情况，首先进入 Advanced Query Form 界面，然后在 Enter Search Terms 检索对话框中输入：

```
phenylketonuria AND genetic
```

也可以在 Global Logic 区内选择 AND，将两个检索词连接起来进行检索。目前的 CRISP 网络版还不允许在一个检索式中使用两个以上的布尔运算符。

在 Expansion Logic 选择 Stem 选项进行截词检索。如 genetic*，检索结果中将包括：genetic、genetics、genetically 等。

点击 Submit Query 执行检索命令。

CRISP 数据库有受控词表，但目前网络版 CRISP 还没有该项功能。可以利用检索结果记录的 Thesaurus Terms 区选最佳检索词，然后返回到检索界面，从而得到更准确的检索结果。

CRISP 检索结果的显示内容有：题目、项目资助号以及项目负责人的姓。点击题目可以看到详细内容，包括文摘。

还可以用主要负责人姓名和研究机构来检索。

28.3.2 National Science Foundation(国家科学基金)

NSF 检索系统用 Verity 检索语言来检索基金资助信息。在 NSF 的检索界面点击 Verity search language 即可进入检索状态。

28.3.2.1 Documents Online(在线文件)

NSF Documents Online(`http://www.nsf.gov/cgi-bin/pubsys/browser/odbrowse.pl`)可以检索到所有 NSF 的联机文件，包括 NSF Program Announcements and Information。在 Documents Online 界面，用户可以在 Full Text 窗口进行基本检索，也可在检索界面中选择 Fielded Search 或 Text Search 可以进行字段检索和高级检索。

在字段检索窗口中用户从 Document Type 选择 Program Announcements and Information 进行检索，相应的检索式出现在 Full Text 窗口。允许使用布尔运算符

和截词符。例如，在 Full Text 窗口输入检索词 genetic*，然后点击位于屏幕底部的 Document Status 的 Current，只检索当前数据库内容，最后点击 Search 执行检索命令，屏幕出现最新的有关遗传研究的信息。用户可以与任一记录链接，获取详细信息。

28.3.2.2 Search Awards(资助检索)

NSF Search Awards(http://www/nsf.gov/verity/srchawd.htm)主页能检索 NSF 提供的资助情况。它有基本检索和字段检索功能(http://www.nsf.gov/verity/srchawdf.htm)，其中 Fielded Search(字段检索)提供的信息比 Documents Online 详细。布尔检索与位于 Full Text 屏幕下部，使用方法与 Documents Online 相同。可以再次执行下面的检索式：

phenylketonuria AND genetic*

点击 Search 按钮执行搜索，检索结果显示在屏幕上，列出资助项目清单，还可以进一步获得资助的详细信息。

Fielded Search 表格供用户从不同字段入手检索，包括调查者(investigator)。如果一个检索策略式需在多个检索字段内检索，则使用位于检索屏幕下部的 Boolean operator used to combine field expressions 功能，选择 AND 运算符。这样可以保证 AND 操作符用于从采用不同的字段得到合并的检索结果。

附录

本章涉及的网站：

Highwire Press: http://highwire.stanford.edu/

Adobe Corporation(Adobe Acrobat Reader for reading PDF files): http://www.adobe.com/

National Institutes of Health Grants and Funding Resources: http://www.nih.gov/grants/

National Science Foundation Grants and Awards Resources: http://www.nsf.gov/home/grants.htm

(逄大欣　译)

参　考　文　献

[1] National Institutes of Health, Office of Extramural Research (December 3, 1997). NIH Guide: Using the TOC Notification LISTSERV Service, National Institutes of Health [1 page]. http://www.nih.gov/grants/guide/listserv.htm.